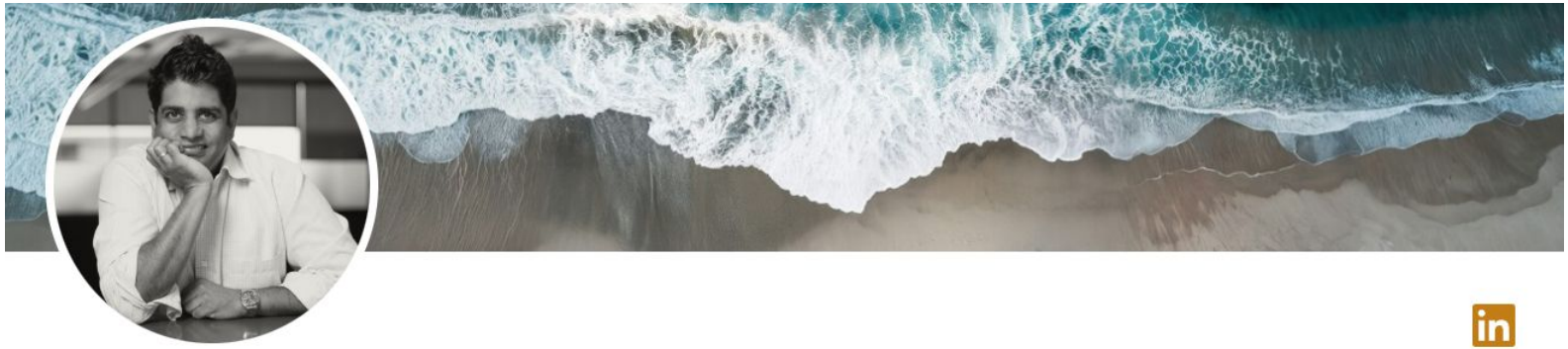
O'REILLY®

# Fundamentals of Large Language Models

# About me



## Jonathan A. Fernandes

AI/ML Engineer Building & Shipping Production-ready GenAI & Large Language Model Solutions Since Before ChatGPT.

United Kingdom · Contact info

jonfernandes

University of Warwick - Warwick Business School

# This online training is always being updated.

- We'll cover the controversy surrounding distillation of Anthropic models by Chinese companies

## What does GPT stand for?

 Generative Pre-trained Transformer

 General Pre-trained Transformer

What are the parameters for a Large Language Model?

 The size of the model

 The variables that get adjusted during the training

# What are tokens for a Large Language Model?

 These are subwords

 This is another word for parameter

## What is the size of the GPT-5.1 model?

 175B

 This information wasn’t released

How can you ensure that GPT-5.1 won't hallucinate?

 Give it more training data

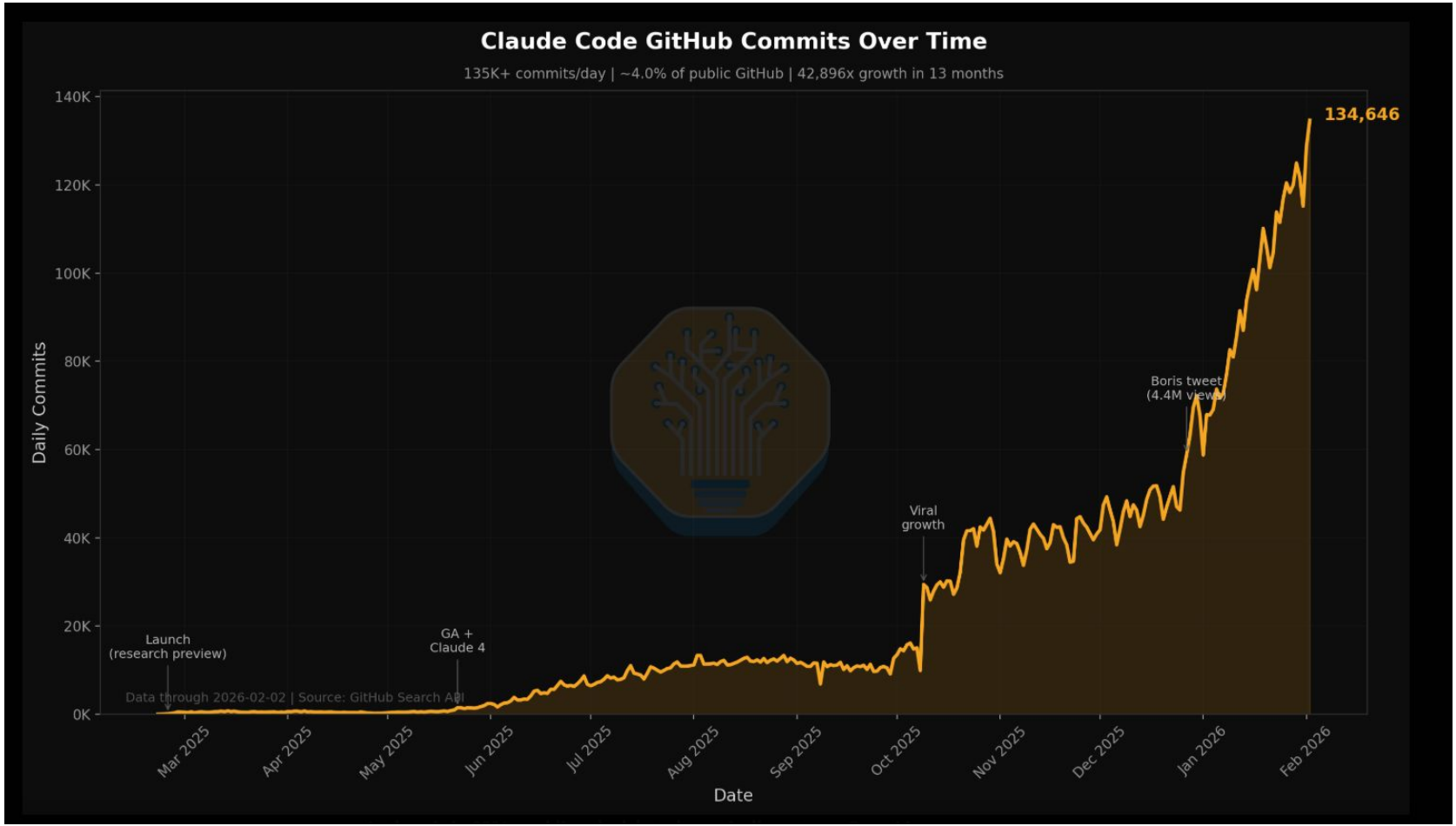
 No way to do this at the moment.

# Is DeepSeek-R1 an open source model?

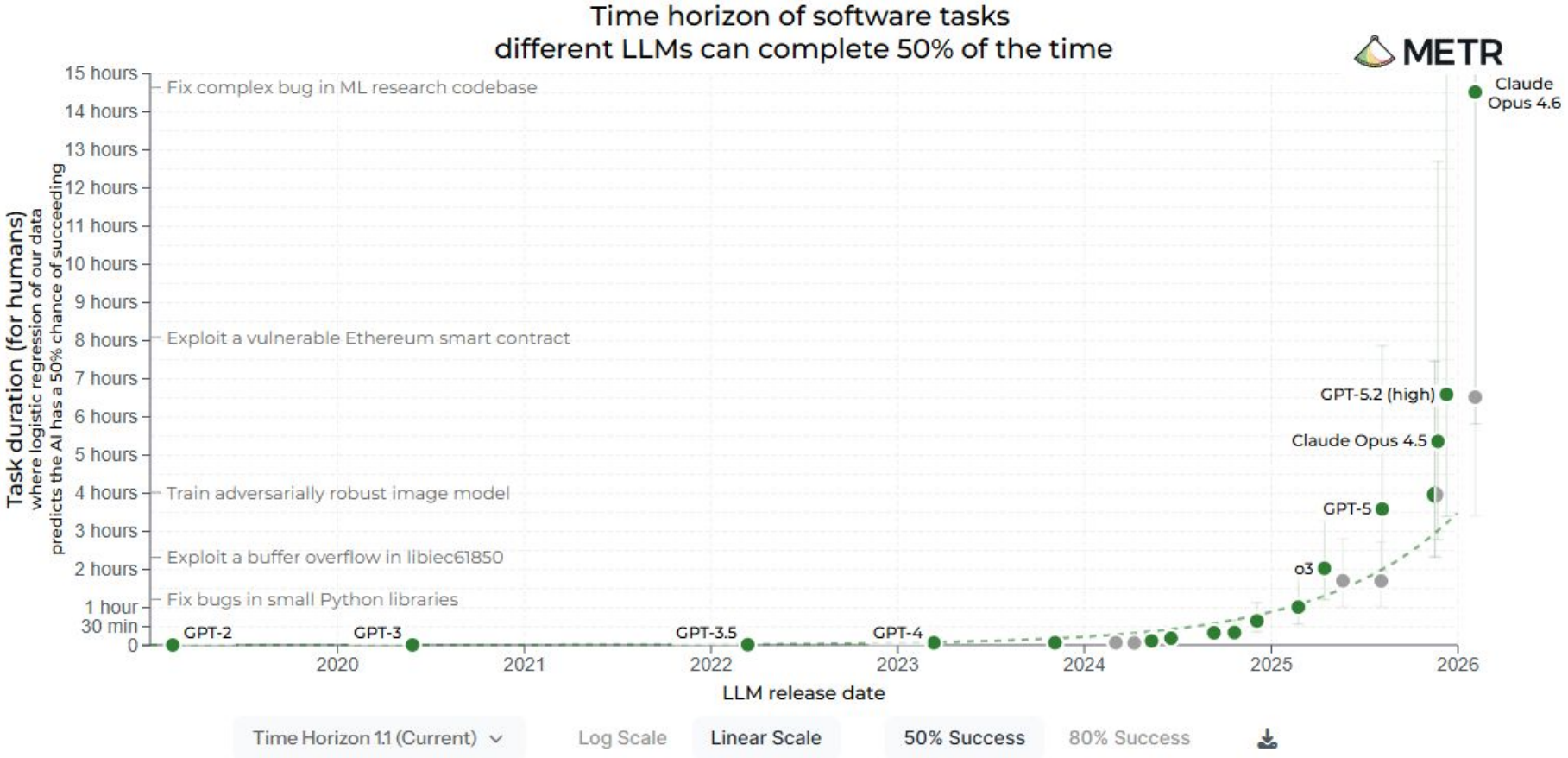
 Yes

 No

# December 2025

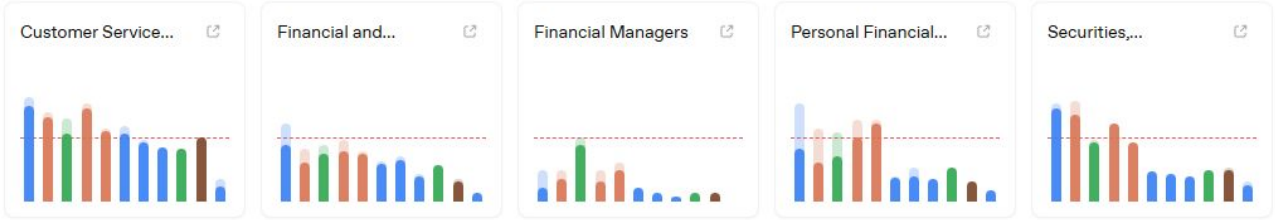



https://newsletter.semianalysis.com/p/claude-code-is-the-inflection-point




https://metr.org/time-horizons/

FINANCE AND INSURANCE

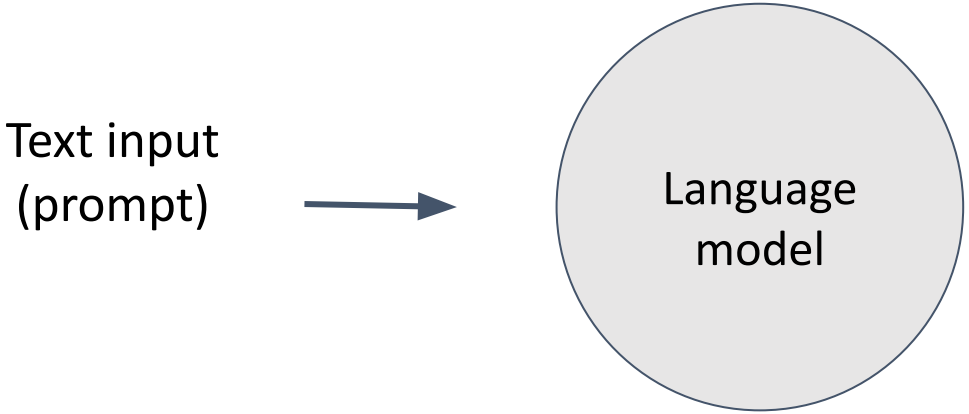


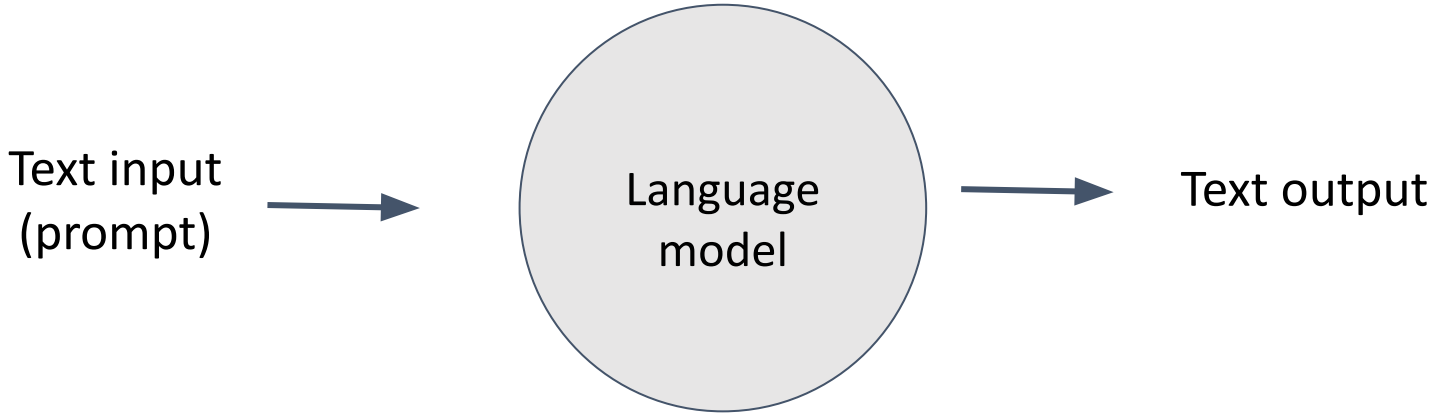

GOVERNMENT

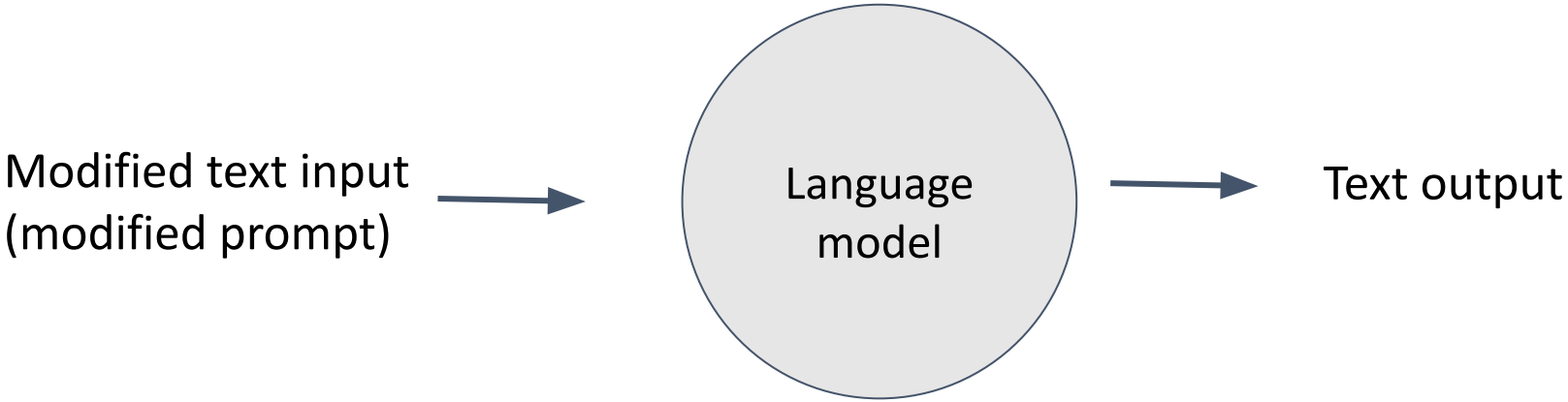

HEALTH CARE AND SOCIAL ASSISTANCE

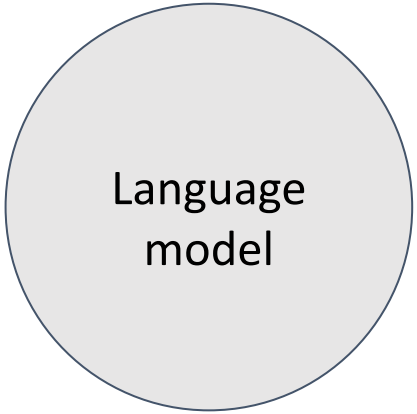

https://evals.openai.com/gdpval/leaderboard

# What Are Large Language Models and GPT?


Text input
(prompt)
Language
model

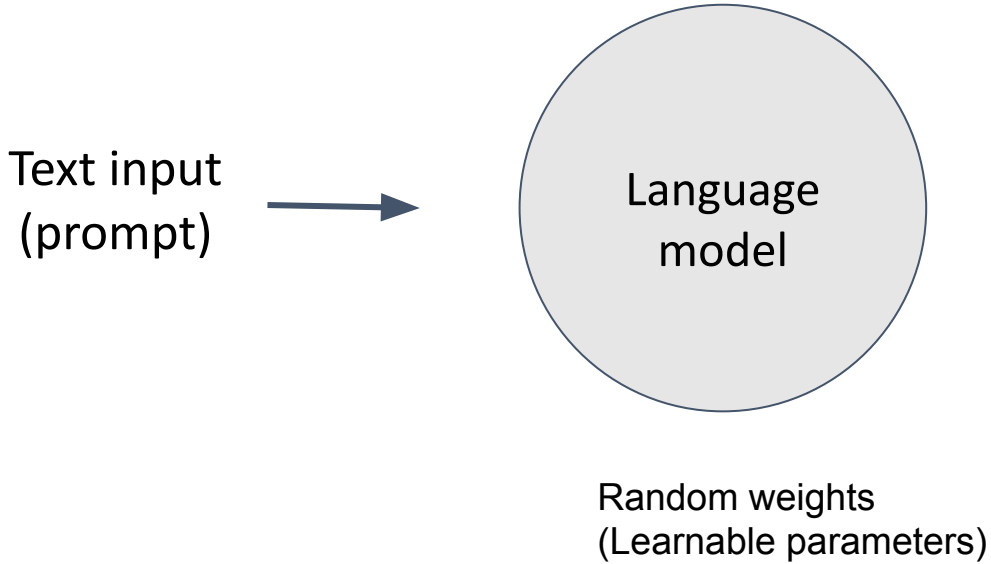

Text input
(prompt)
Language
model
Text output


Modified text input
(modified prompt)
Language
model
Text output

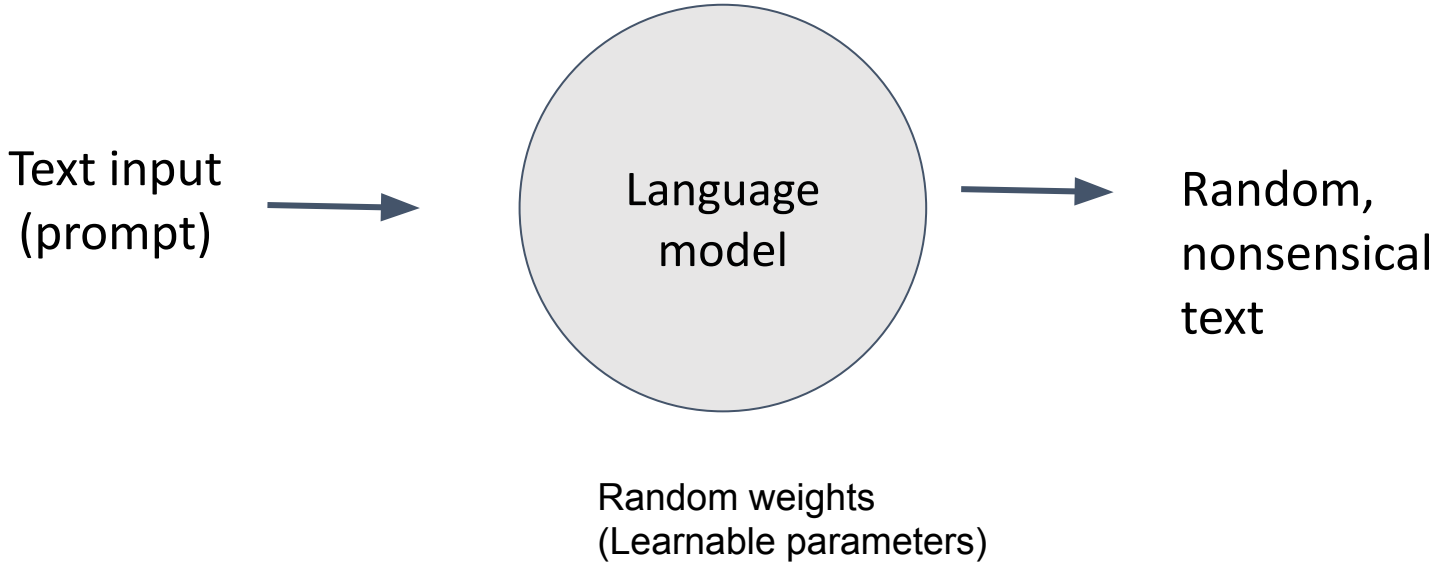



Parameters or
Learnable parameters

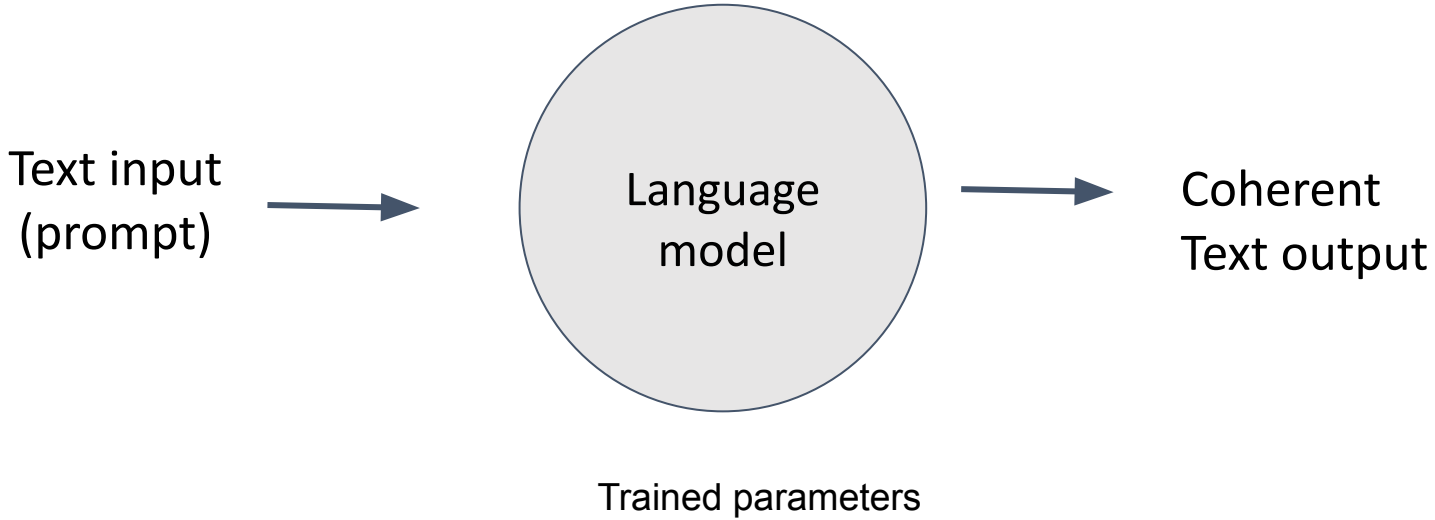
# Initial model (Before training Language model)



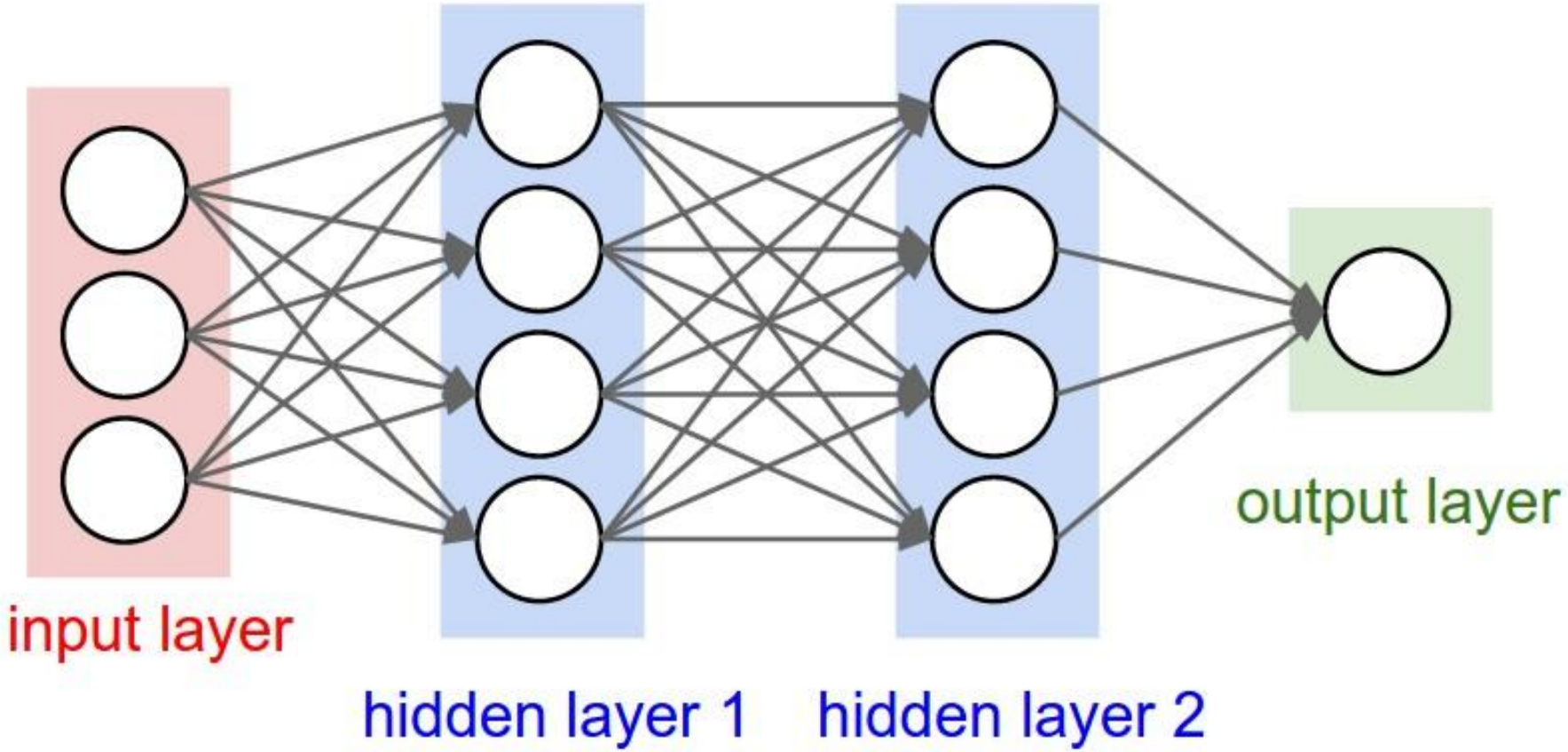
# Initial model (Before training Language model)

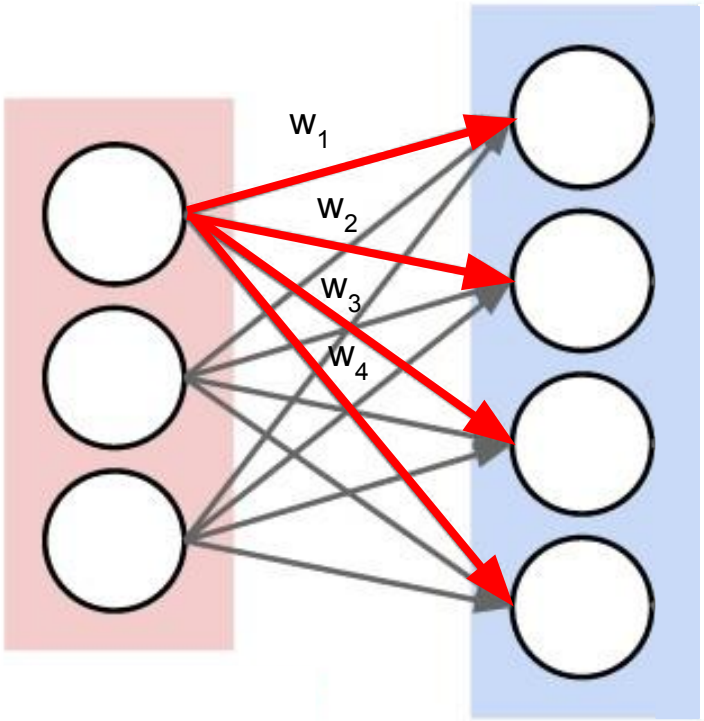


After training



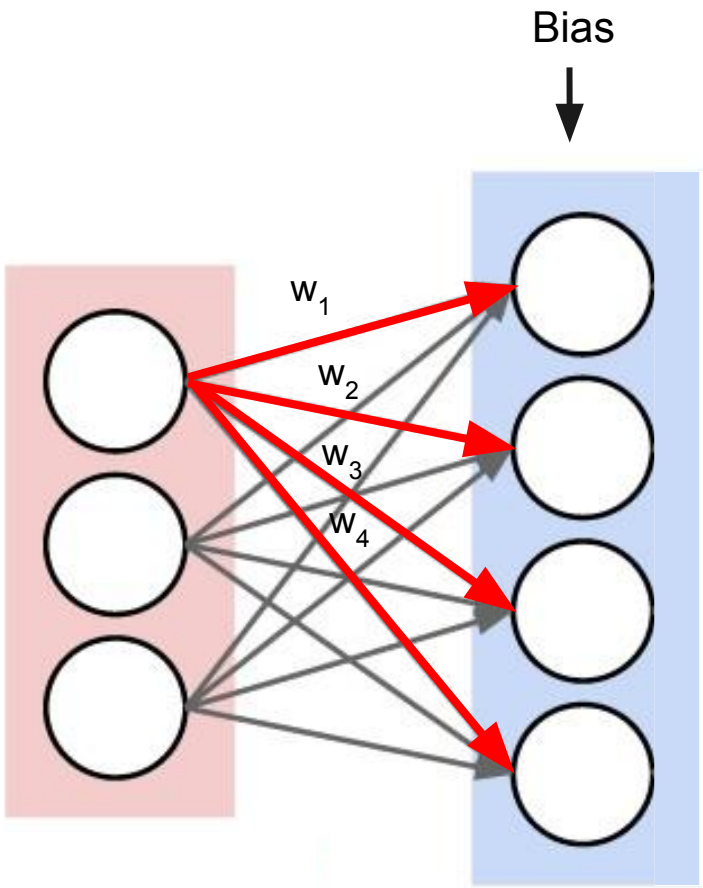
# What are these (learnable) parameters?


input layer
hidden layer 1
hidden layer 2
output layer

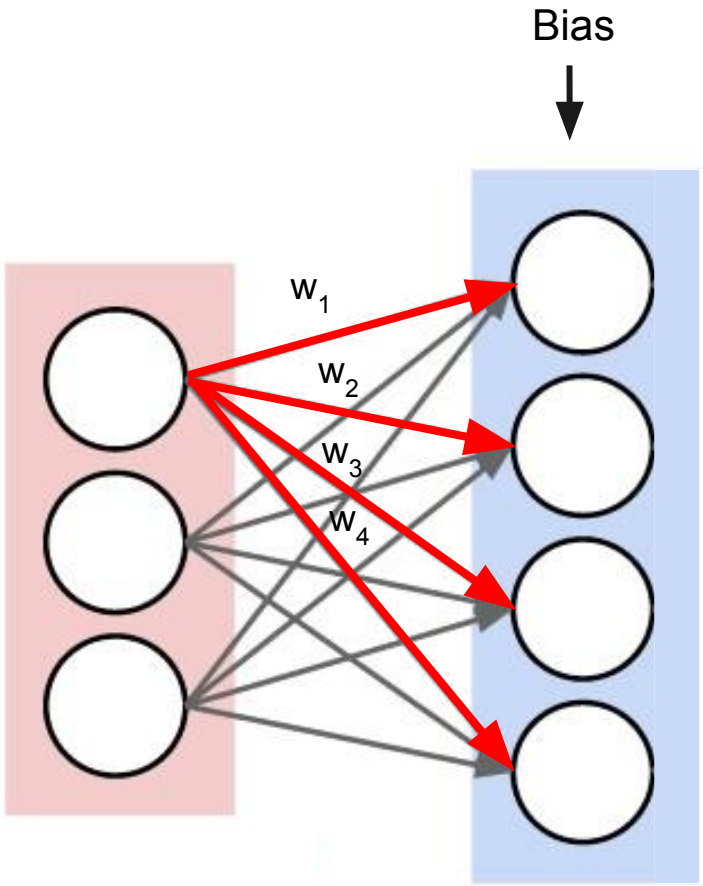



# Weights:

## Inputs x Outputs



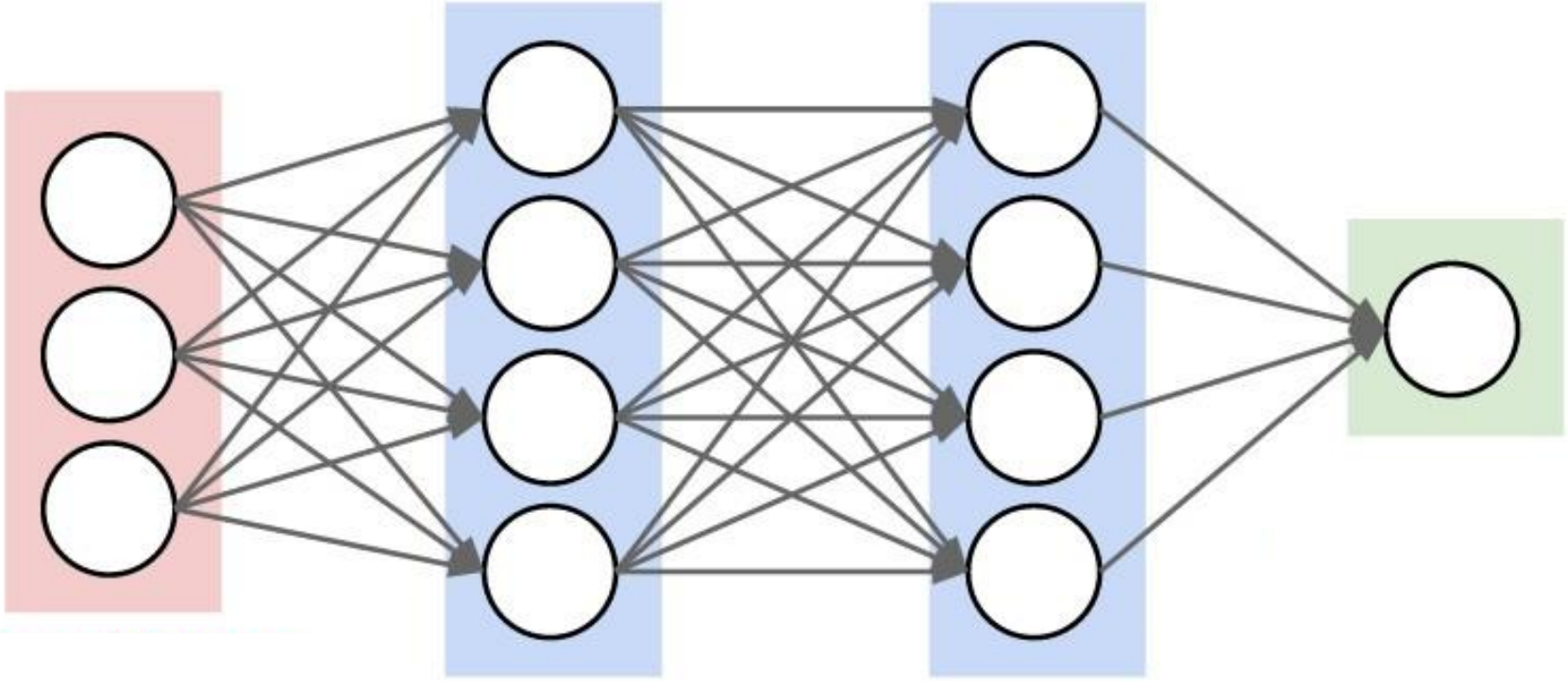

# Bias




Number of parameters:
Number of weights + Biases
(3 x 4) + 4

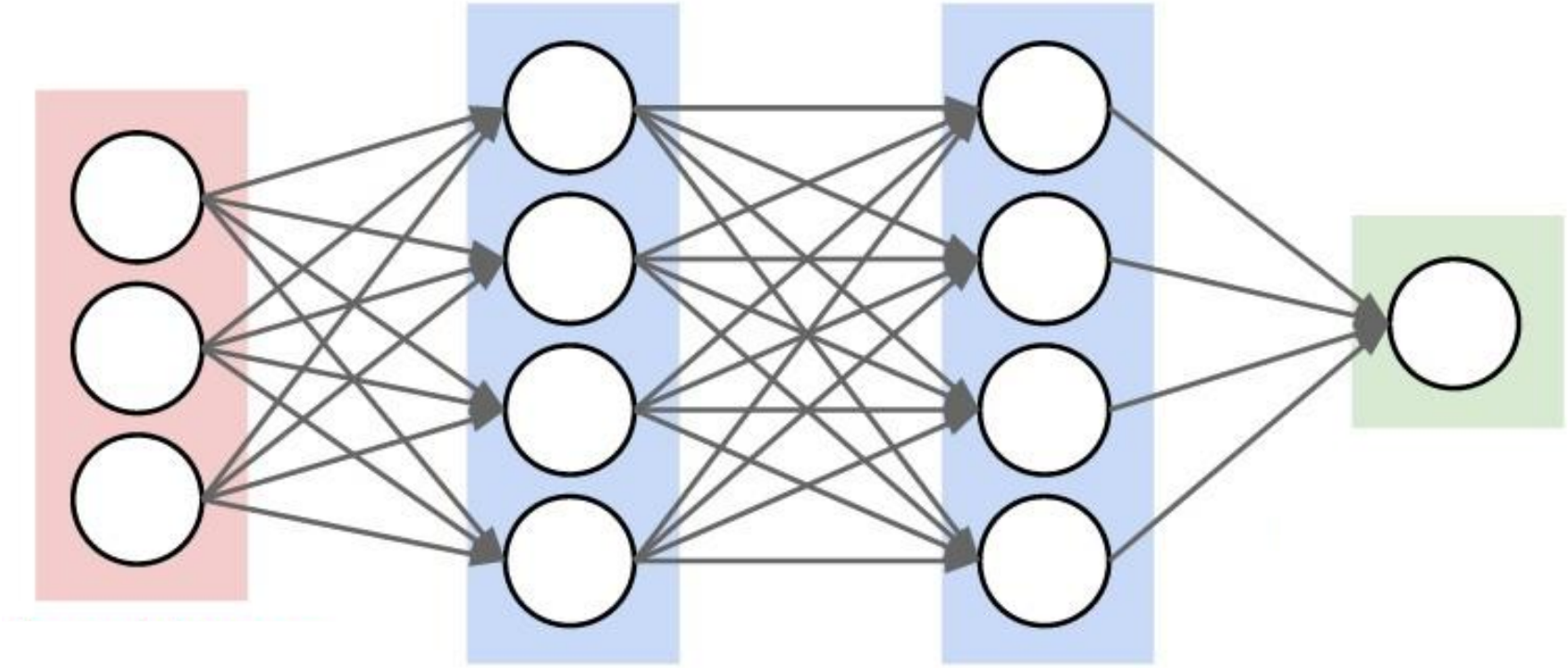
3x4 + 4

**16**



3x4 + 4
**16**

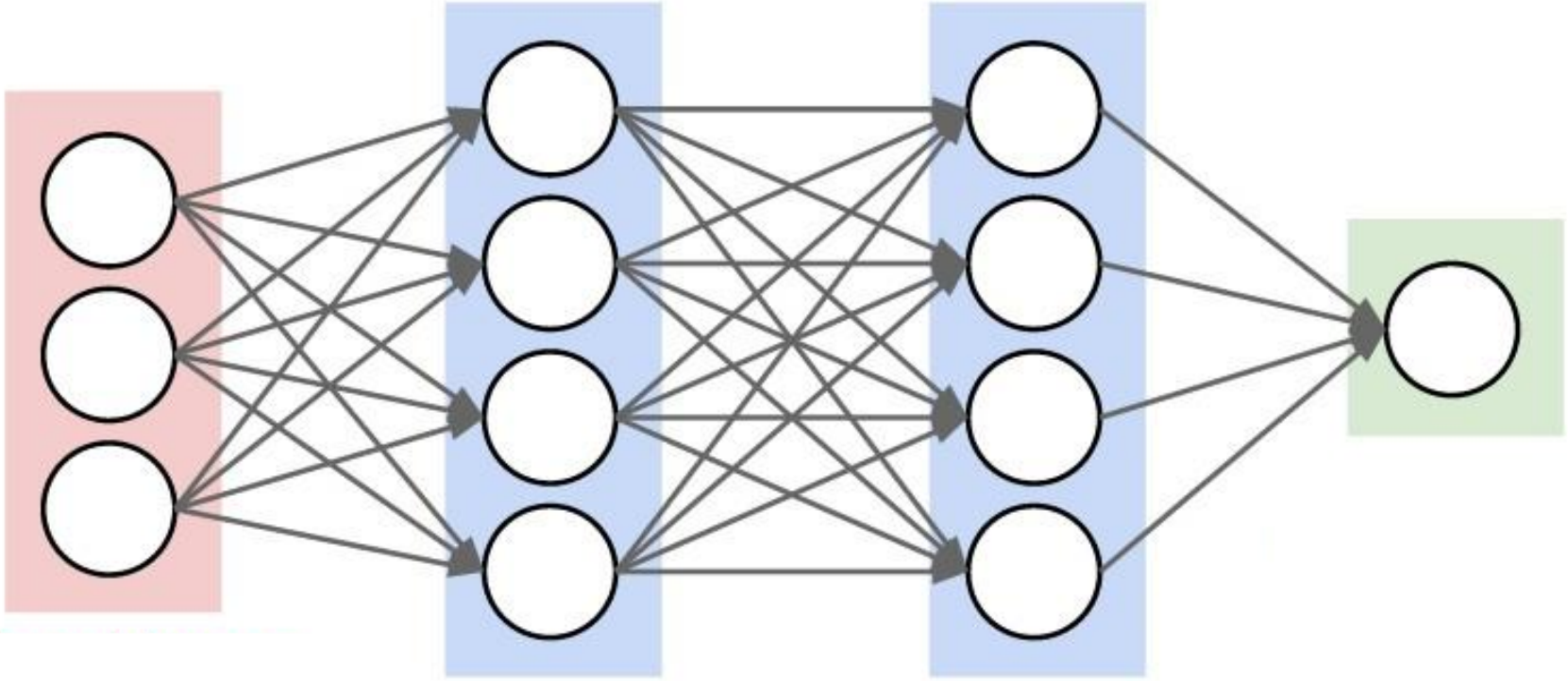
4x4 + 4
**20**



3x4 + 4
**16**

4x4 + 4
**20**

4x1 + 1
**5**

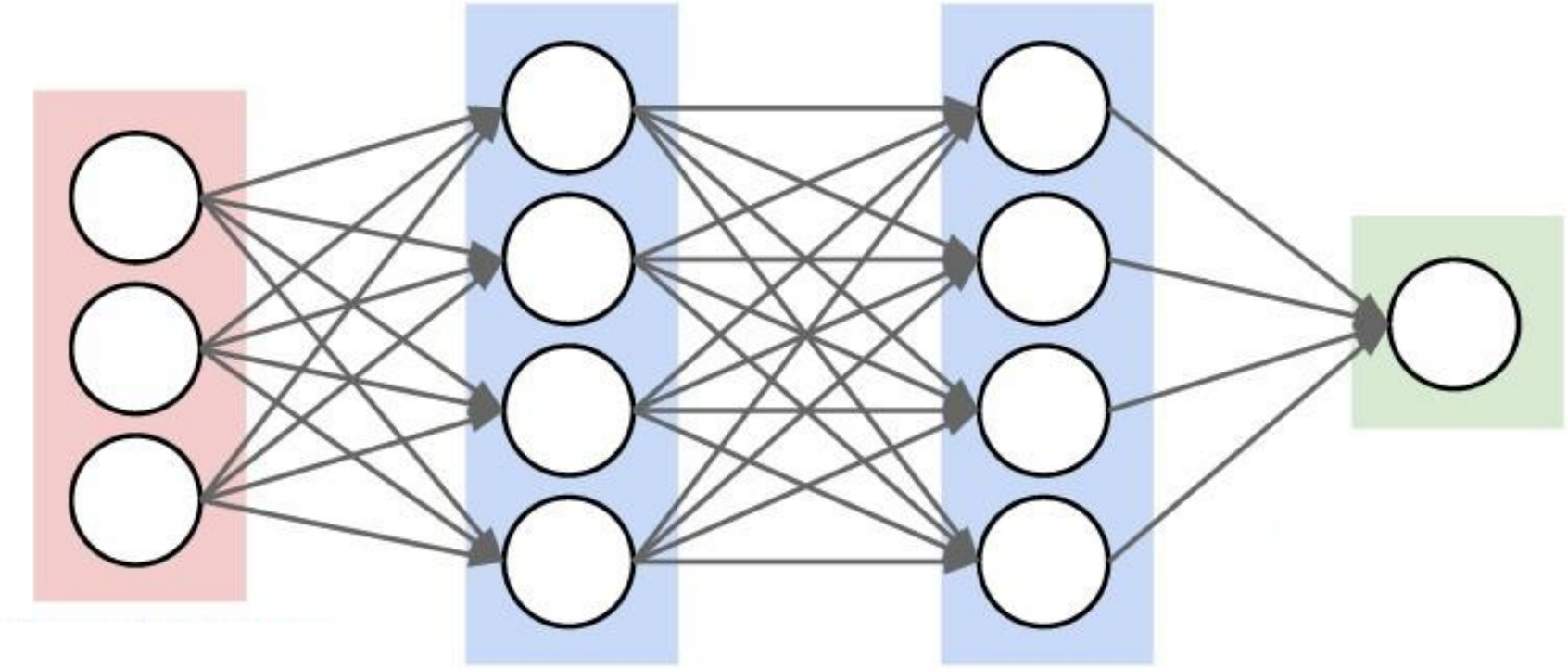


3x4 + 4
**16**

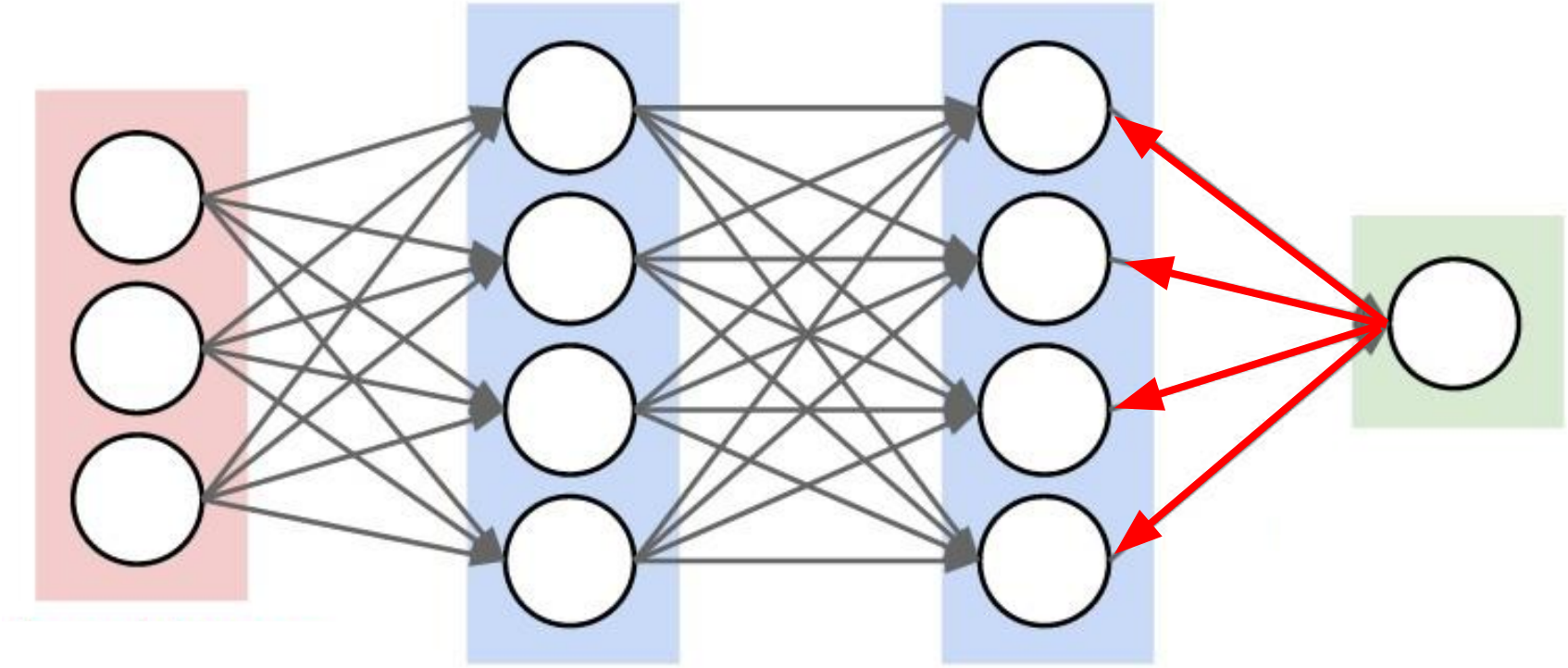
4x4 + 4
**20**

4x1 + 1
**5**

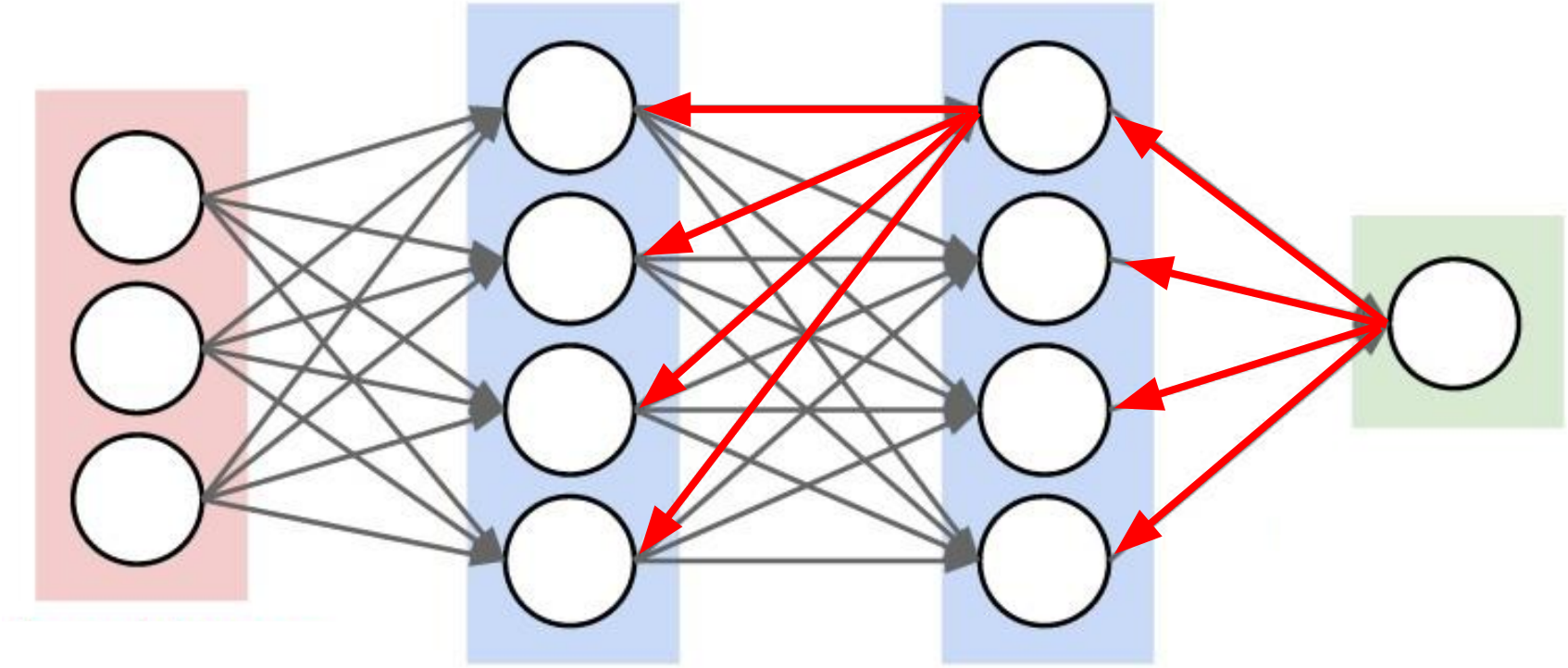
Total parameters:
**41**



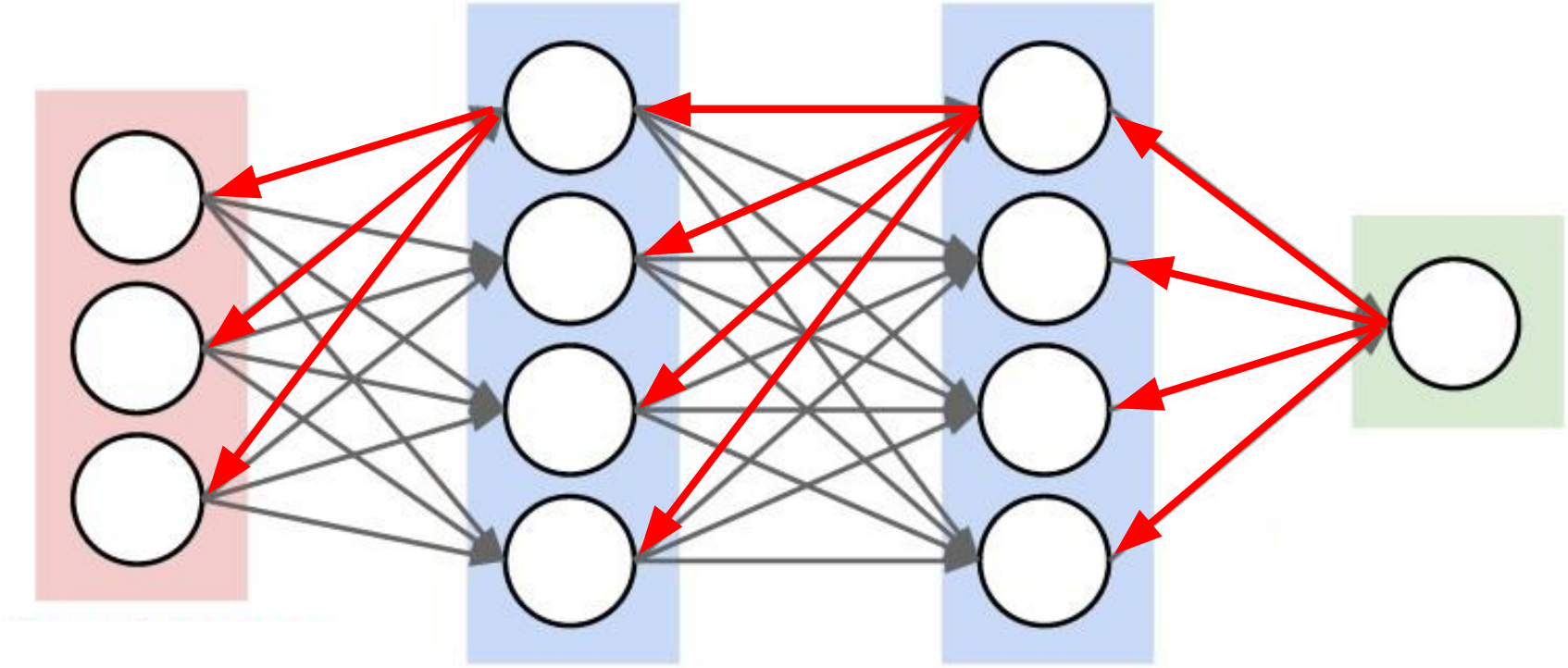
**Backpropagation**



Backpropagation

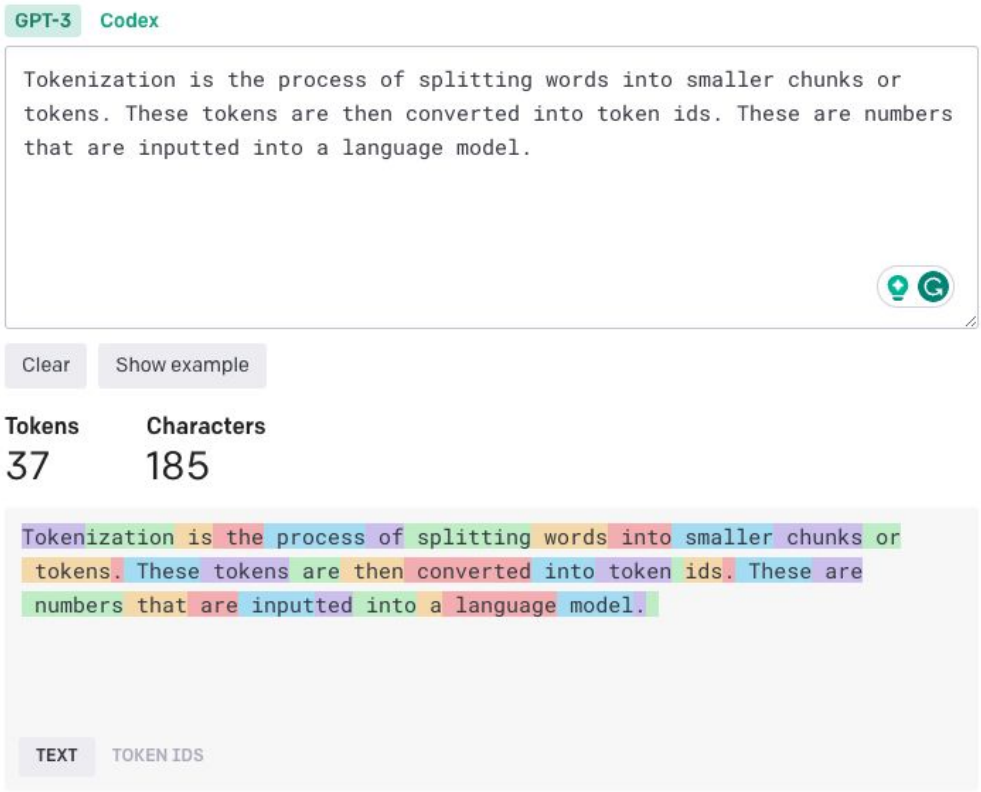


**Backpropagation**



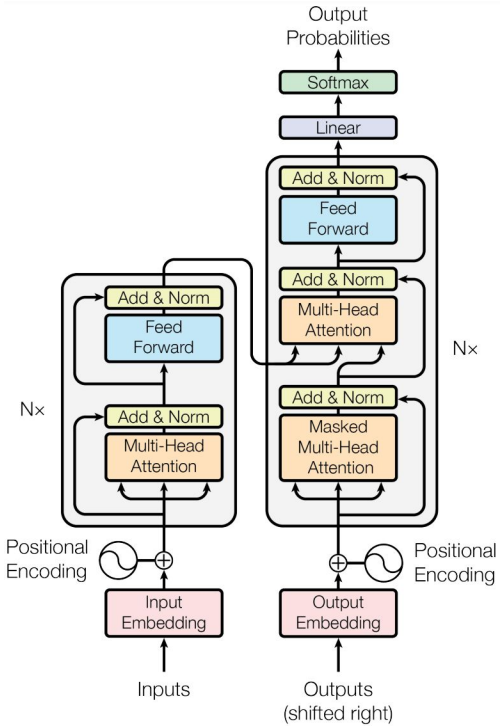
# What are tokens?

# What are tokens?

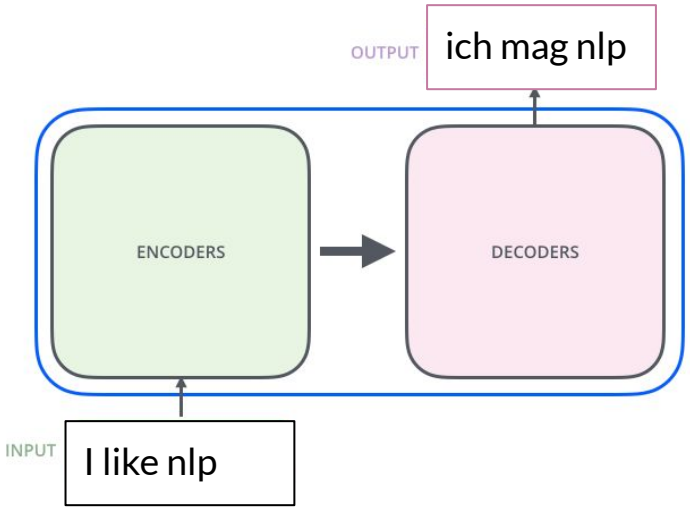


# Transformer: Architecture Overview

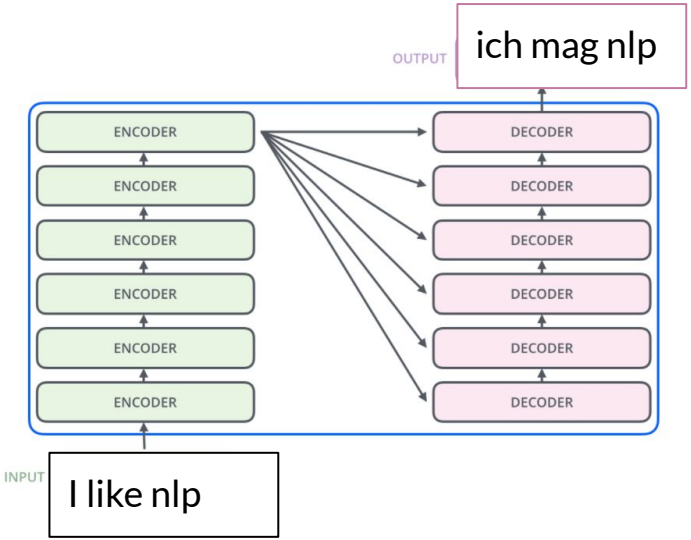
# Transformer architecture



# Transformer overview

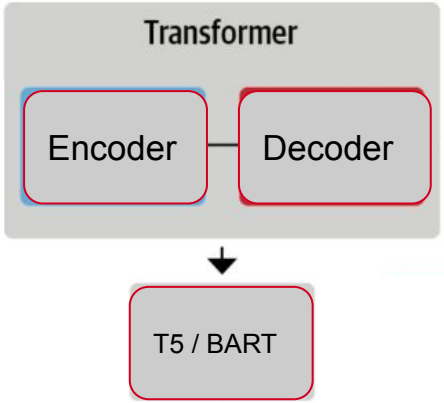


# Transformer overview



# Encoder-decoder model

- Generative tasks



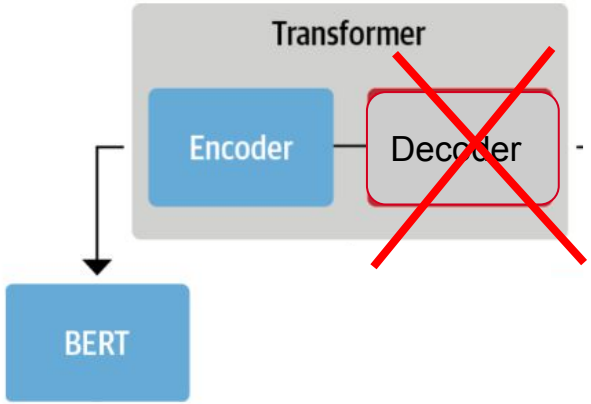
# Encoder-only model

Understanding of input

- Sentence classification
- Named Entity Recognition

Family of BERT models:

- BERT, RoBERTa, DistilBERT ...

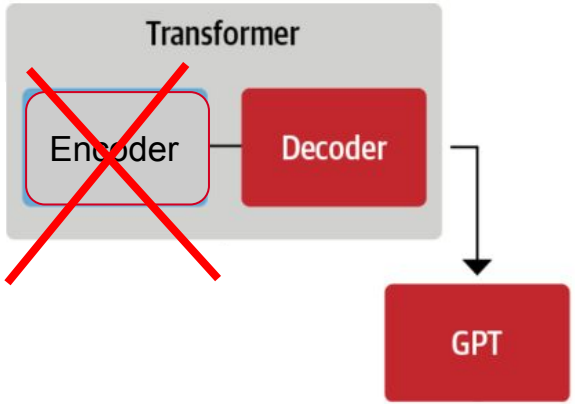


# Decoder-only model

- Generative tasks

Examples:

- OpenAI GPT models, Claude, Gemini



**G**enerative

**P**re-trained

**T**ransformer

**G**enerative – predicting a future token, given past tokens

**P**re-trained

**T**ransformer

**G**enerative – predicting a future token, given past tokens

**P**re-trained – trained on a large corpus of data

**T**ransformer

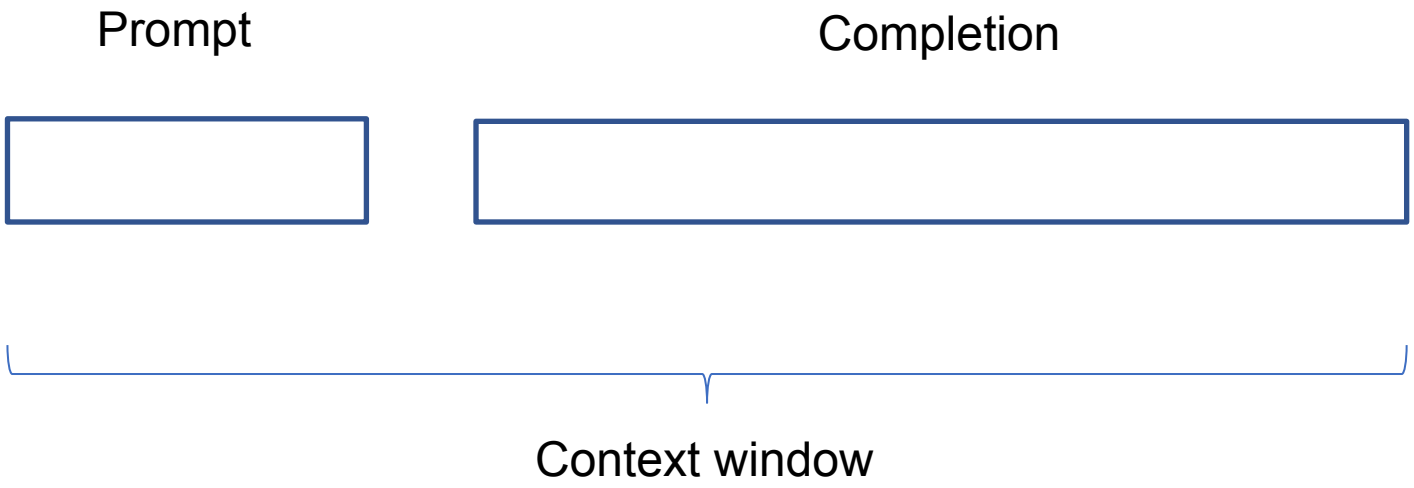
**G**enerative – predicting a future token, given past tokens

**P**re-trained – trained on a large corpus of data

**T**ransformer – portion of transformer architecture

# Context length

# Context window / length



# Differences between models

gpt-4: 8,000 tokens
gpt-4-32k: 32,000 tokens

gpt-3: 2,000 tokens
gpt-3.5: 4,000 tokens
gpt-4-turbo 128,000 tokens

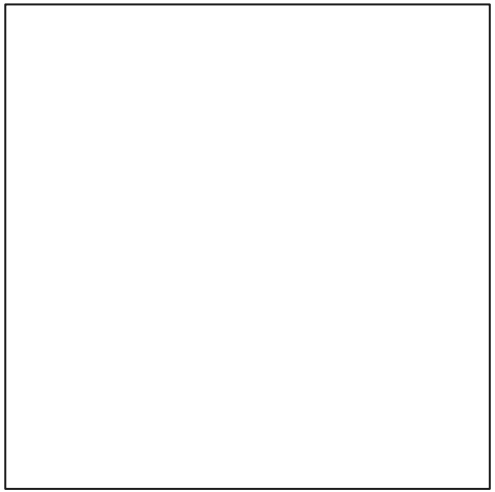
claude-2 100,000 tokens
claude-3 200,000 tokens

Gemini Pro: 1mil+ tokens
Gemini Pro 1.5: 2mil tokens

# How do you run an LLM?

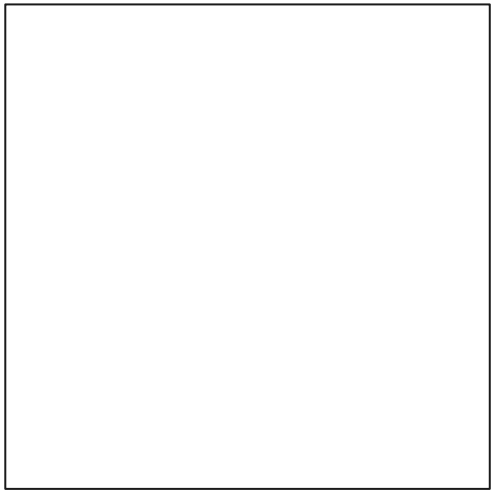
# How do you run an LLM?

File to load the LLM (written in c)

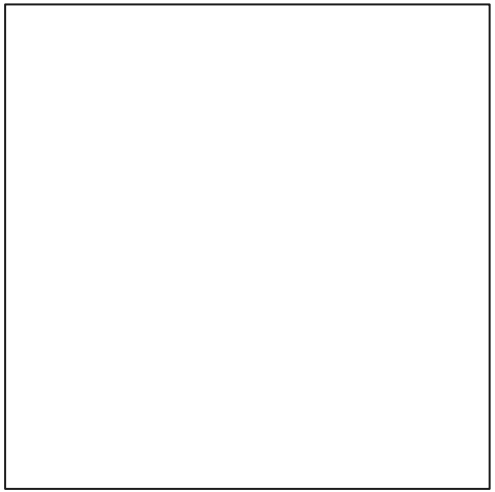


# How do you run an LLM?

File to load the LLM (written in c)

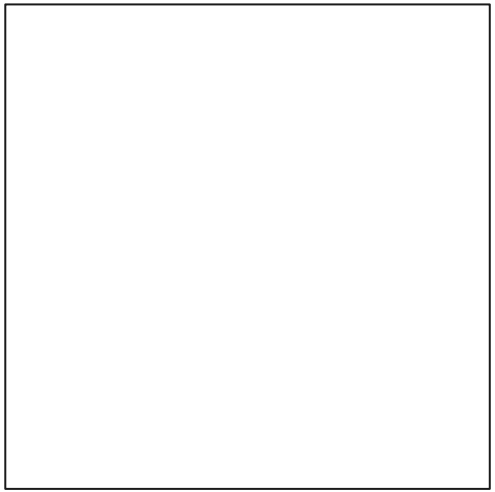


Parameters

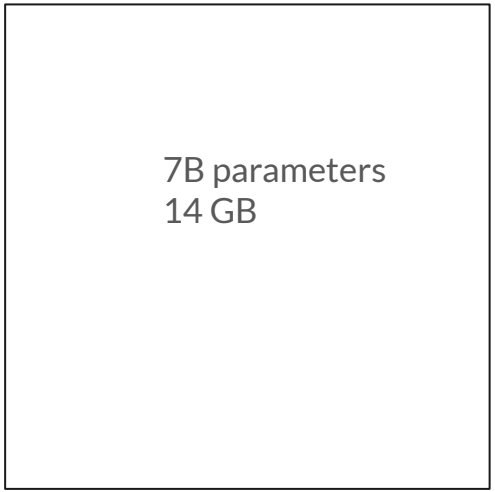


# How do you run an LLM?

File to load the LLM (written in c)



Parameters



# How do you run an LLM?


NVIDIA A100 80GB

# What are the different kinds of prompts?

# What are the different kinds of prompts?

- Prompts
- System prompts

# Objectives of earlier LLMs

- Predict the next word

# Challenges with earlier LLMs

- Doesn't follow user instructions
- Can generate toxic language
- Can make up facts

# Objectives of post-GPT-3.x / GPT-4.x

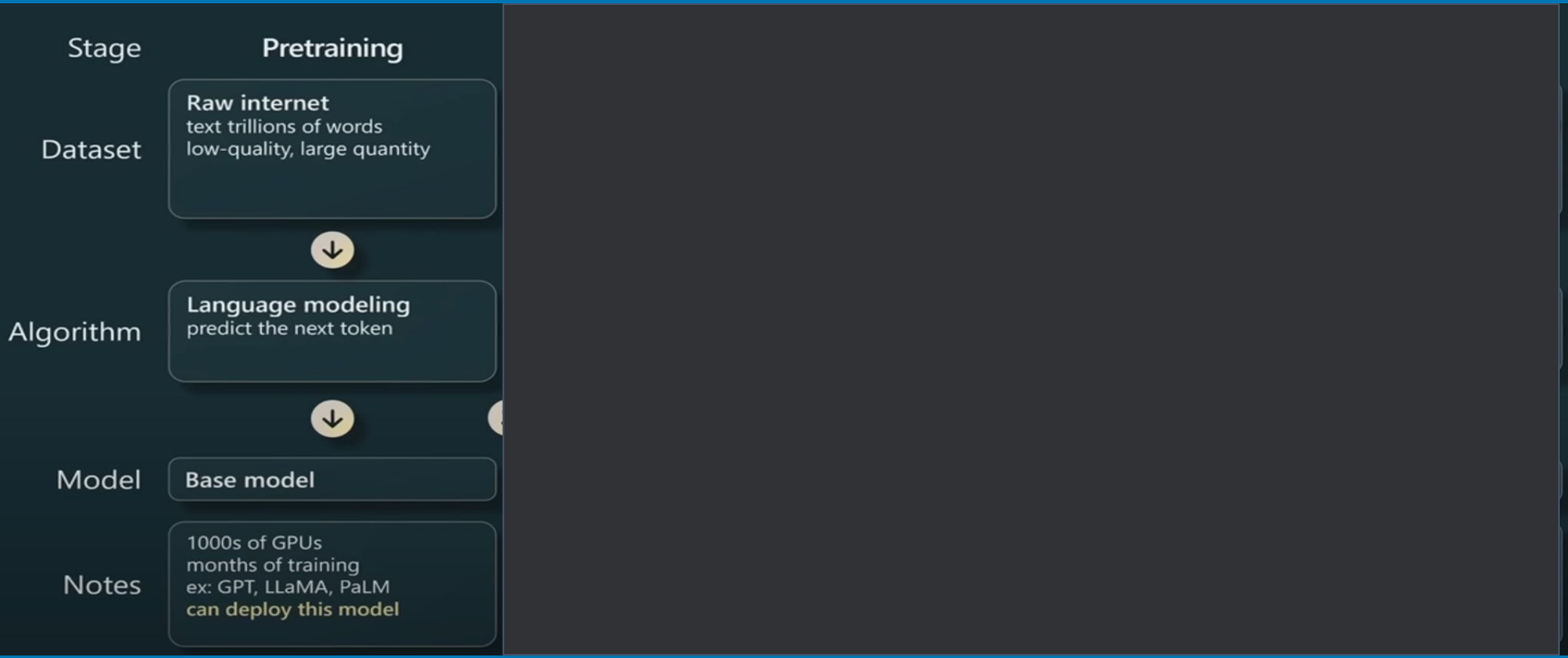
- Helpful and able to follow instructions
- Not toxic – for example, hateful speech, foul language
- Less likely to fabricate information or hallucinate

# OpenAI playground

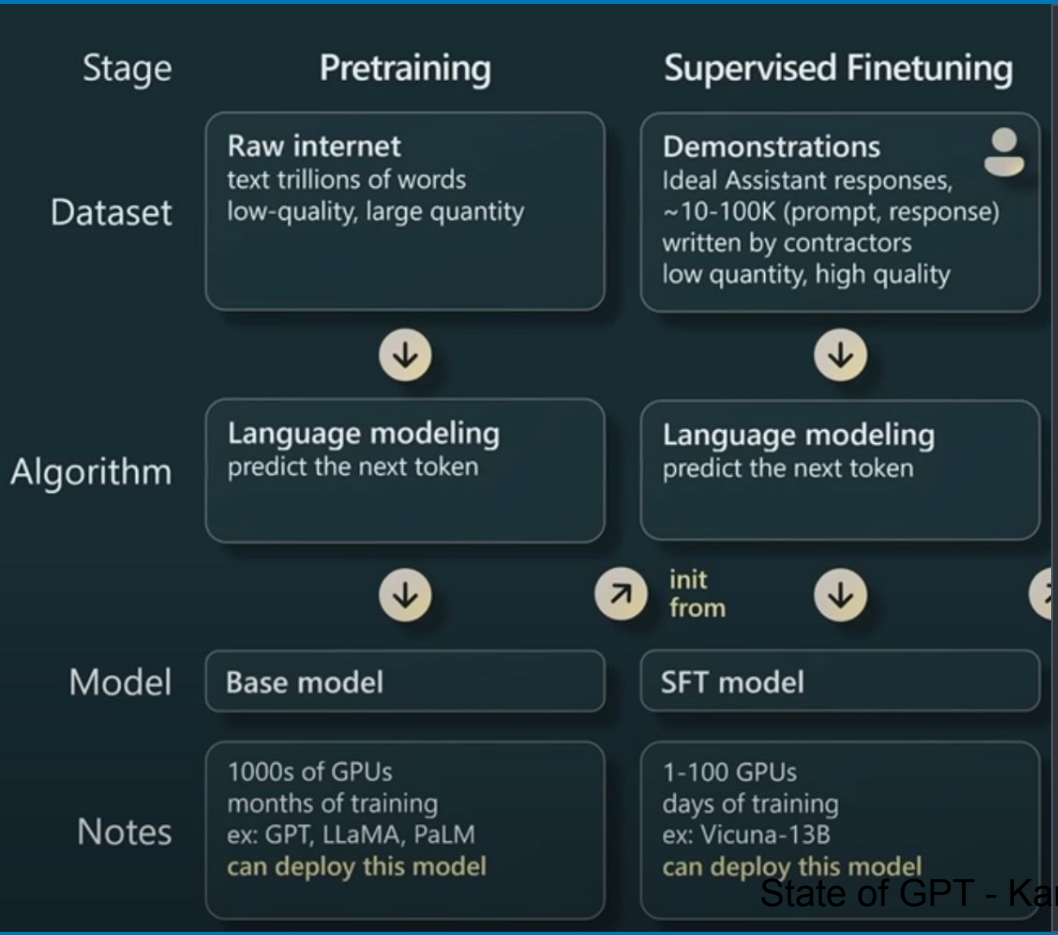
Create a shopping list

# How are LLMs trained?

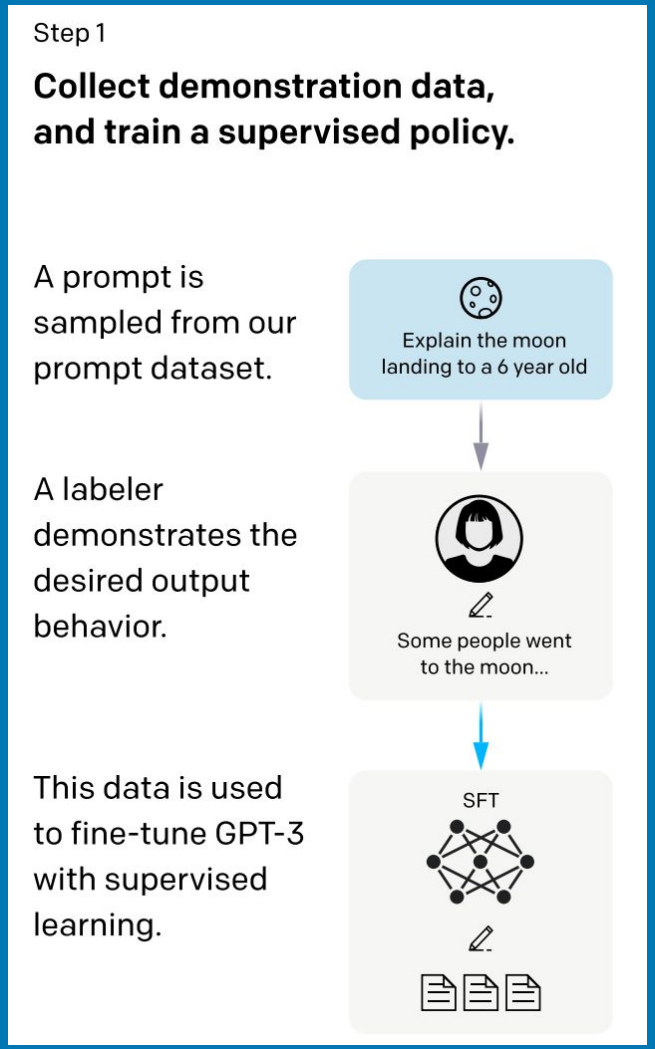
# Pre-training



# Instruction Tuning




State of GPT - Karpathy




Training language models to follow instructions with human feedback - Long et al (Open AI)

# Example: Summarize a News Article

Submit Skip « Page 3 / 11 » Total time: 05:39

## Instruction

Include output

Summarize the following news article:

====
{article}
====

## Output A

summary1

**Rating (1 = worst, 7 = best)**

1 2 3 4 5 6 7

| | | |
|---|---|---|
| *Fails* to follow the correct instruction / task **?** | Yes | No |
| Inappropriate for customer assistant **?** | Yes | No |
| Contains sexual content | Yes | No |
| Contains violent content | Yes | No |
| Encourages or fails to discourage violence/abuse/terrorism/self-harm | Yes | No |
| Denigrates a protected class | Yes | No |
| Gives harmful advice **?** | Yes | No |
| Expresses moral judgment | Yes | No |

**Notes**

(Optional) notes

## Ranking outputs

**To be ranked**

B A team of researchers from Yale University and University of California, Davis studied the vocalization patterns of several different types of parrots. They found that parrots like to mimic human speech, and can produce a wide range of sounds, such as whistles, squawks, and other types of vocalizations...

C Parrots have been found to have the ability to understand numbers. Researchers have found that parrots can understand numbers up to six. In a series of experiments, the parrots were able to identify the amount of food items under a number of cups...

**Rank 1** ***(best)***

A A research group in the United States has found that parrots can imitate human speech with ease, and some of them can even do so in the same way as humans. This group studied the sounds that parrots make in their natural habitats and found that they use their tongues and beaks in ways that are strikingly...

**Rank 2**

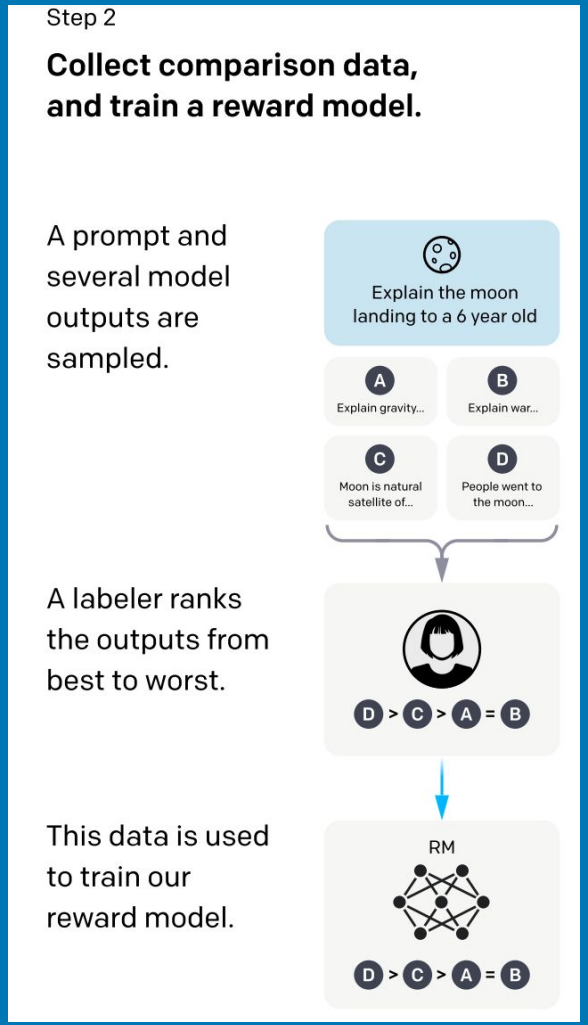
**Rank 3**

E Scientists have found that green-winged parrots can tell the difference between two noises that are the same except for the order in which they are heard. This is important because green-winged parrots are known to imitate sounds. This research shows that they are able to understand the difference between sounds.

D Current research suggests that parrots see and hear things in a different way than humans do. While humans see a rainbow of colors, parrots only see shades of red and green. Parrots can also see ultraviolet light, which is invisible to humans. Many birds have this ability to see ultraviolet light, an ability

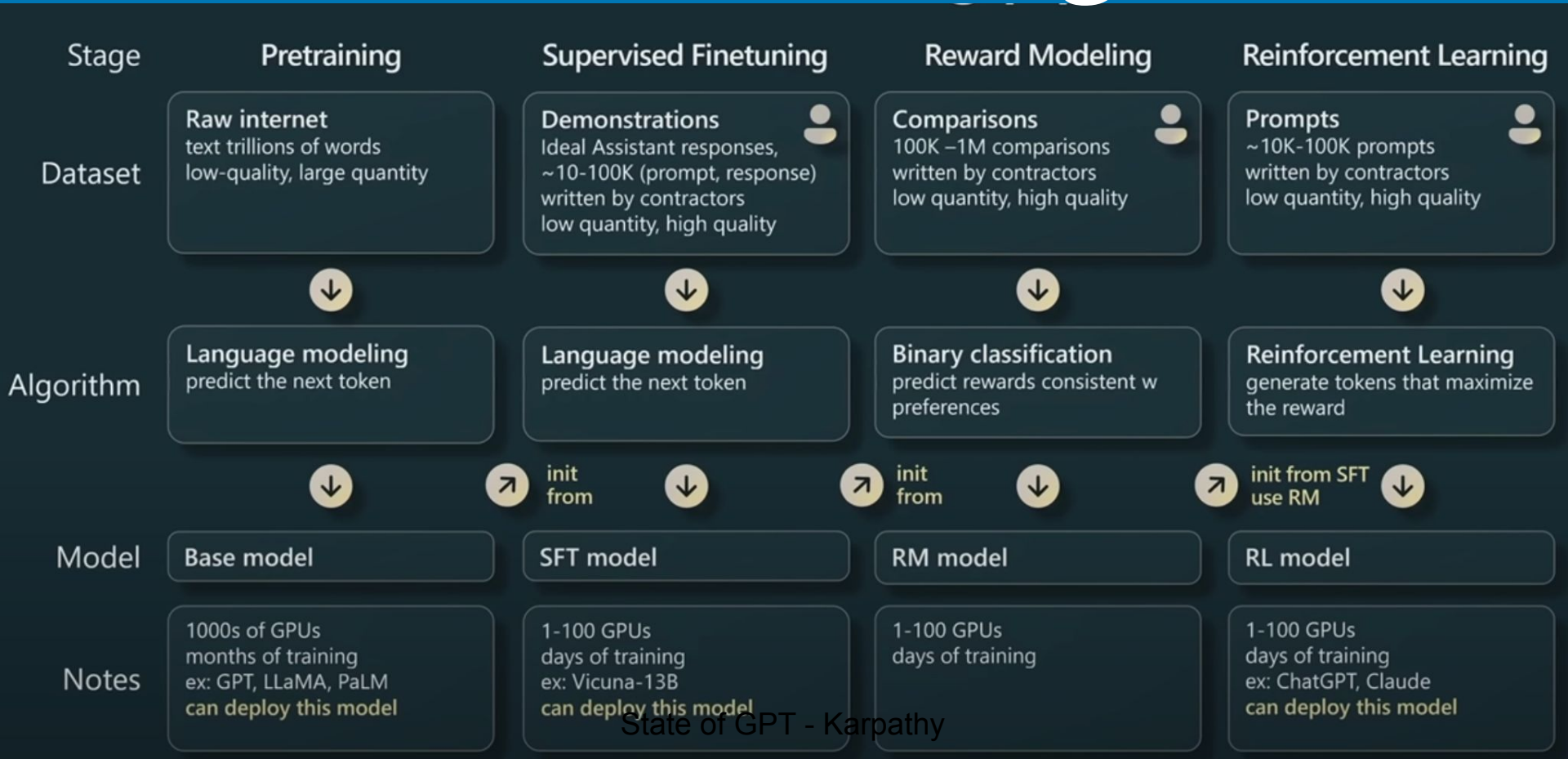
**Rank 4**

**Rank 5** ***(worst)***



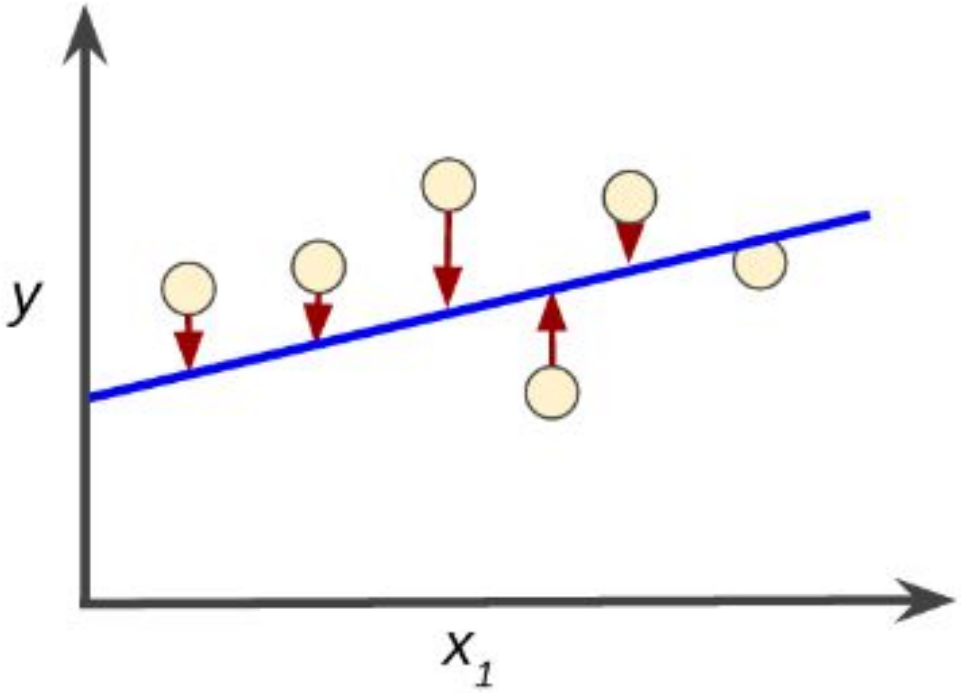

Training language models to follow instructions with human feedback - Long et al (Open AI)

# Preference Tuning

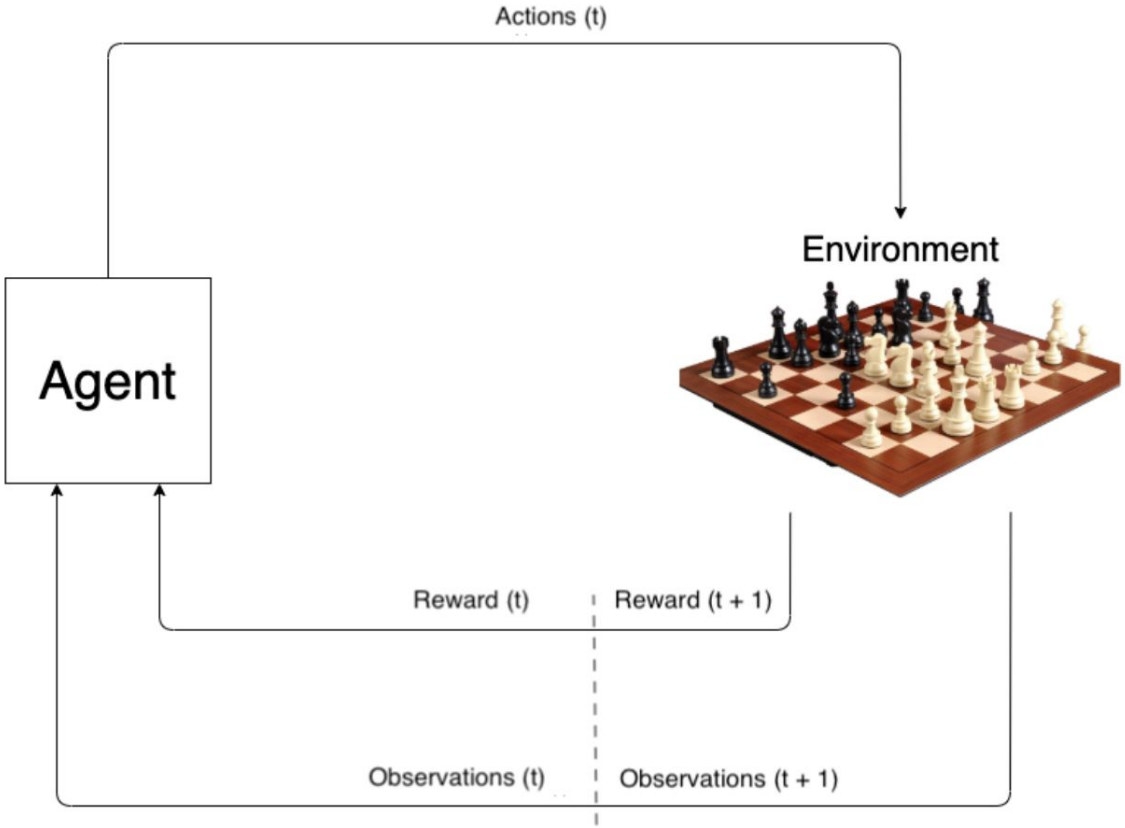



State of GPT - Karpathy

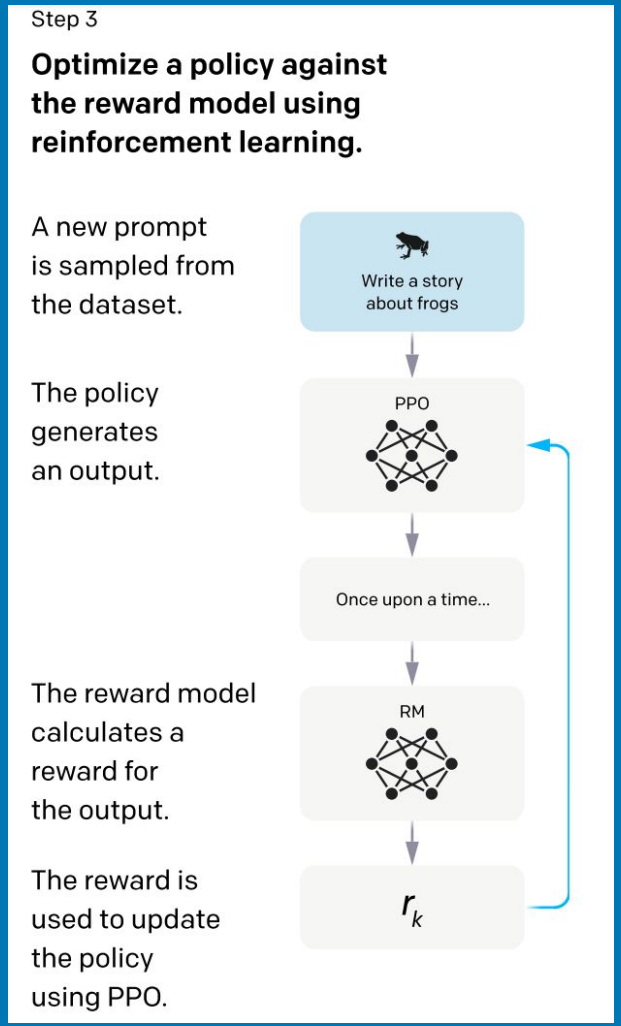
# How ML models are trained



# Reinforcement Learning



**R**einforcement
**L**earning from
**H**uman
**F**eedback




Training language models to follow instructions with human feedback - Long et al (Open AI)

Step 1

**Collect demonstration data, and train a supervised policy.**

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

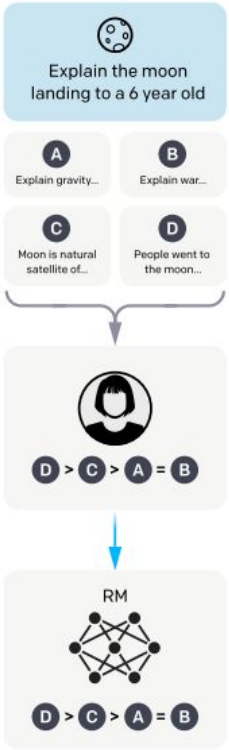

Step 2

**Collect comparison data, and train a reward model.**

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.




Step 3

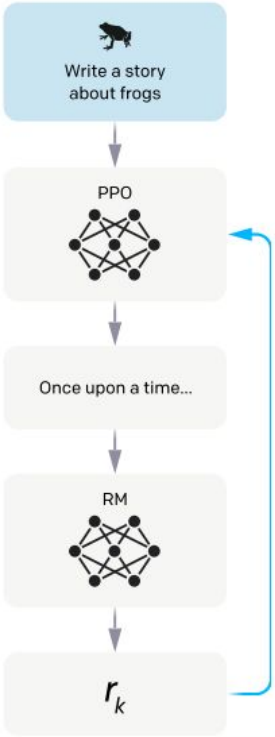
**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



# Updates to how LLMs are trained

# RLHF with **PPO**

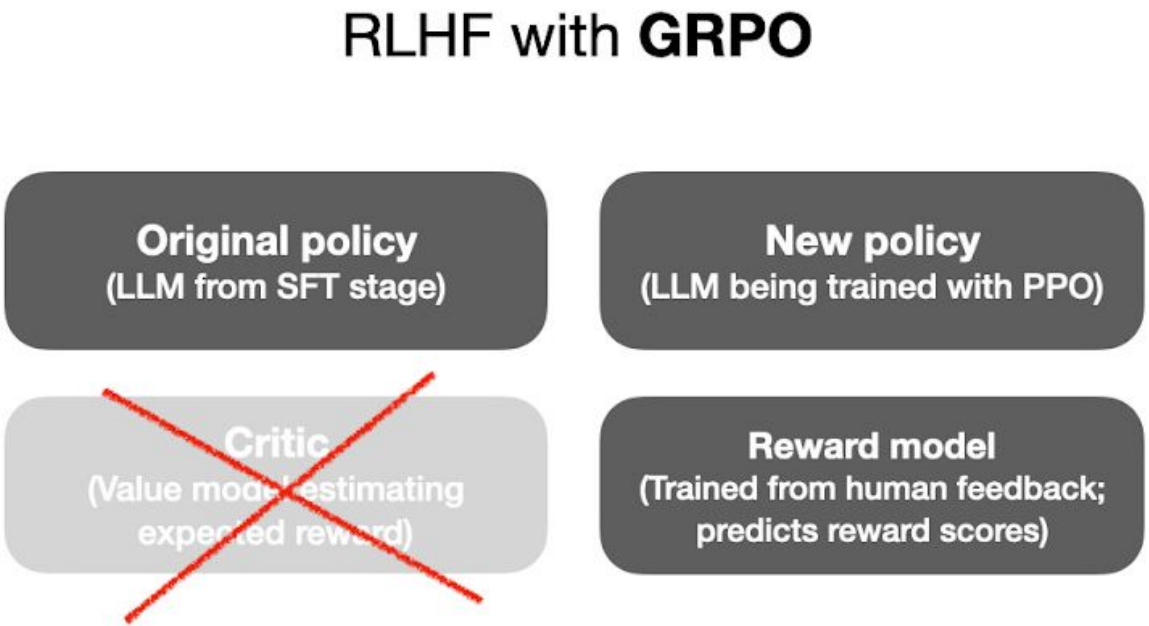
**Original policy**
(LLM from SFT stage)

**New policy**
(LLM being trained with PPO)

**Critic**
(Value model estimating expected reward)

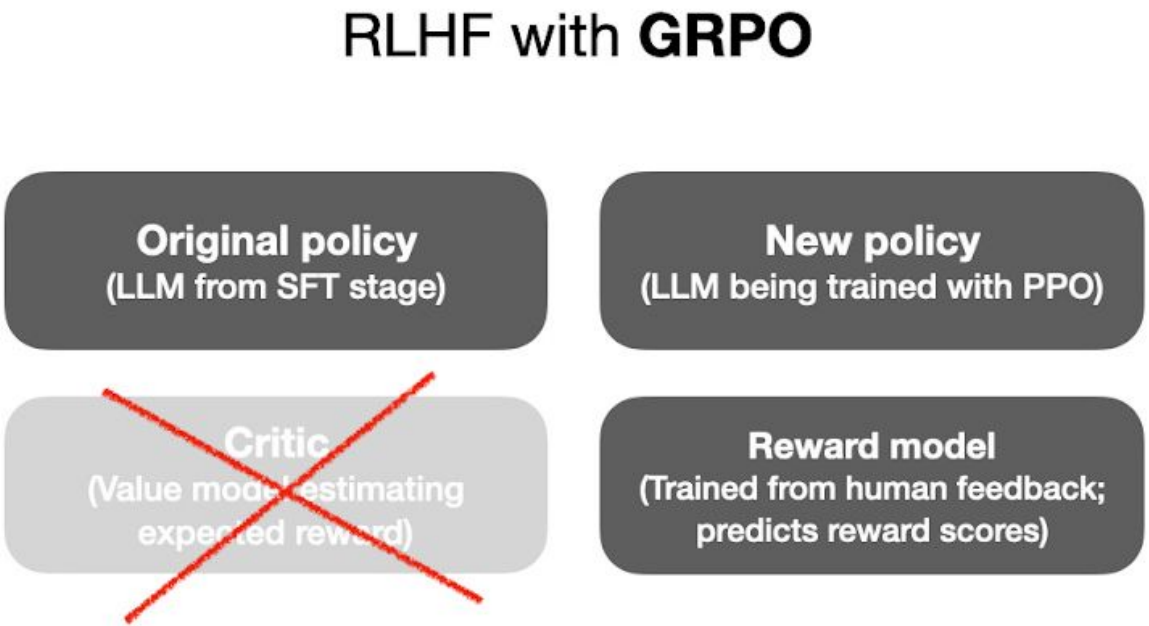
**Reward model**
(Trained from human feedback; predicts reward scores)

# From PPO to GRPO



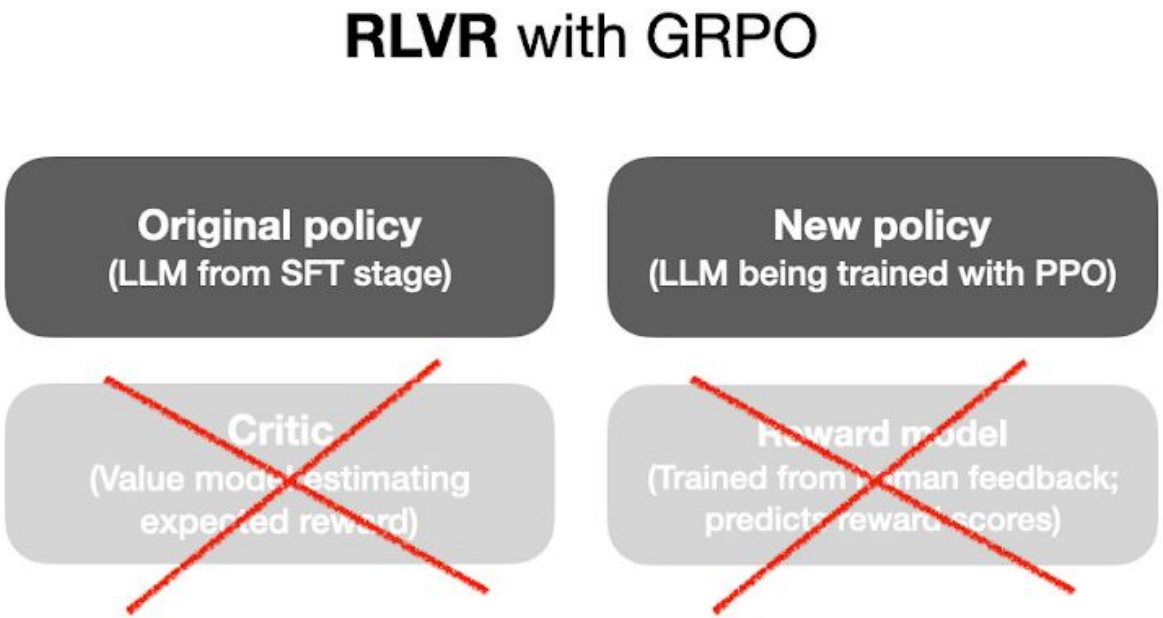

- GRPO introduced by DeepSeek
- A variant of Proximal Policy Optimization (PPO) that enhances mathematical reasoning abilities while concurrently optimizing the memory usage of PPO.

# From PPO to GRPO



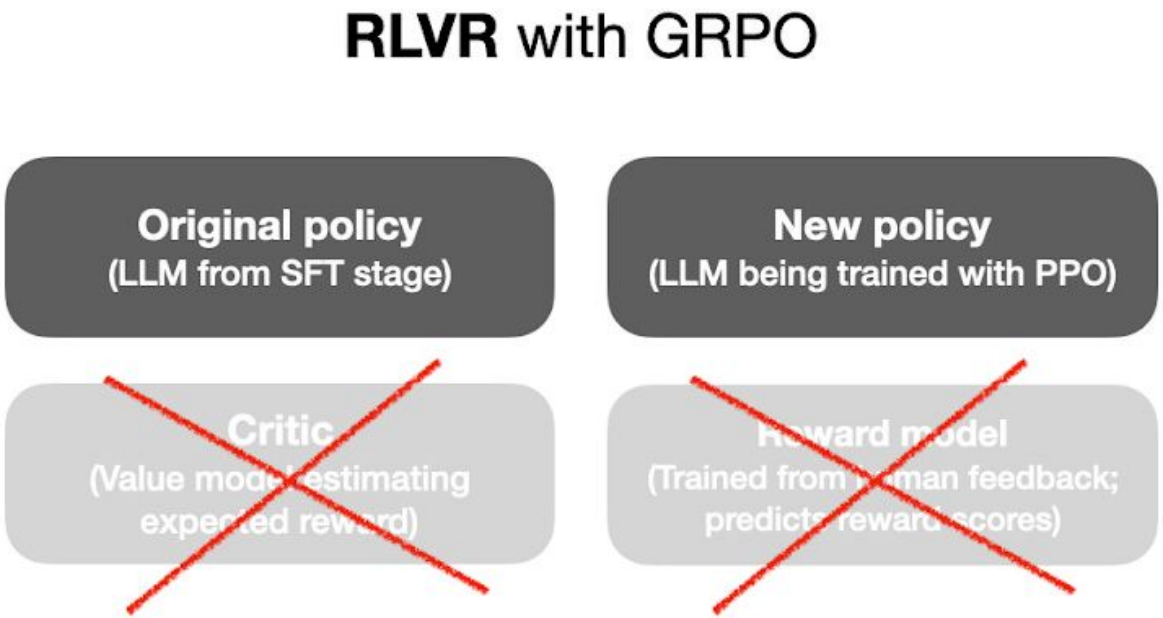

- Drop the "critic" (value model), and  samples multiple answers from the policy model itself and uses their relative quality to compute the advantages.

# From RLHF to RLVR




- Instead of relying on human preferences and training a reward model, the DeepSeek-R1 team used verifiable rewards

# From RLHF to RLVR




- Drop reward model
- Rather than learning what counts as a "good" answer from human-labeled data, model gets direct binary feedback (correct or wrong) from a deterministic tool. Calculators for math problems. Compilers for code generation.

# Scaling and training times

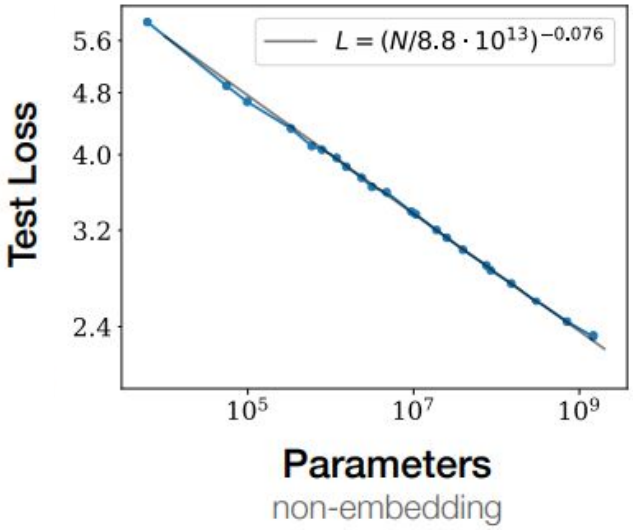
# Scaling Laws

# Scaling Laws

Performance of large models, function of:

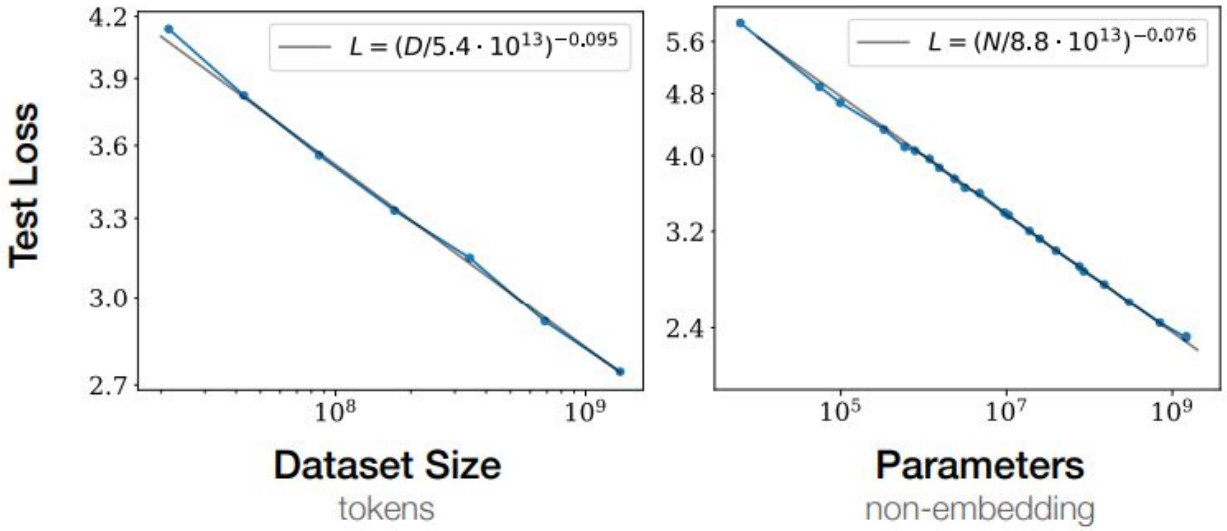
- Model parameters
- Size of the dataset
- Total amount of compute available

# Number of parameters



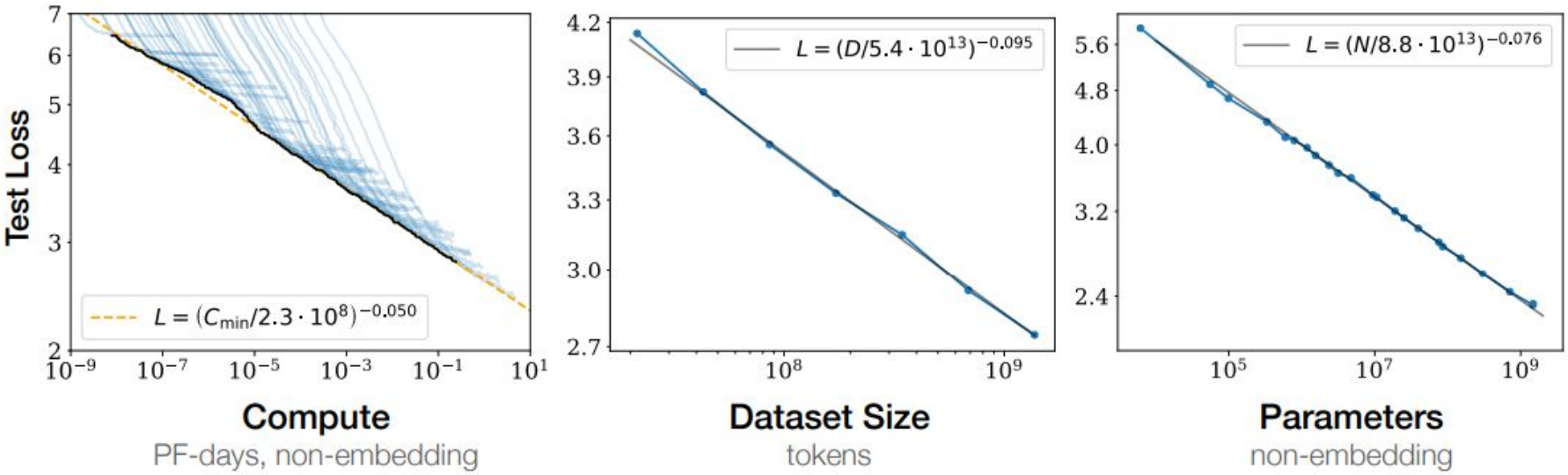

Source: Scaling Laws for Neural Language Models (Kaplan et. al)

# Size of the dataset

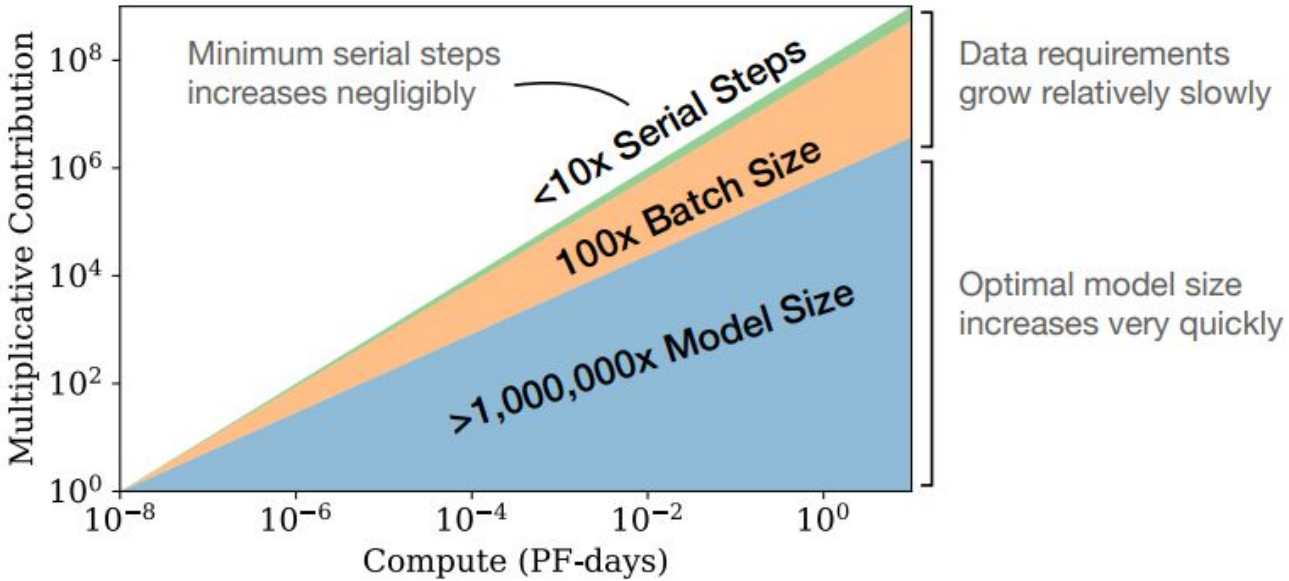



Source: Scaling Laws for Neural Language Models (Kaplan et. al)

# Compute




Source: Scaling Laws for Neural Language Models (Kaplan et. al)



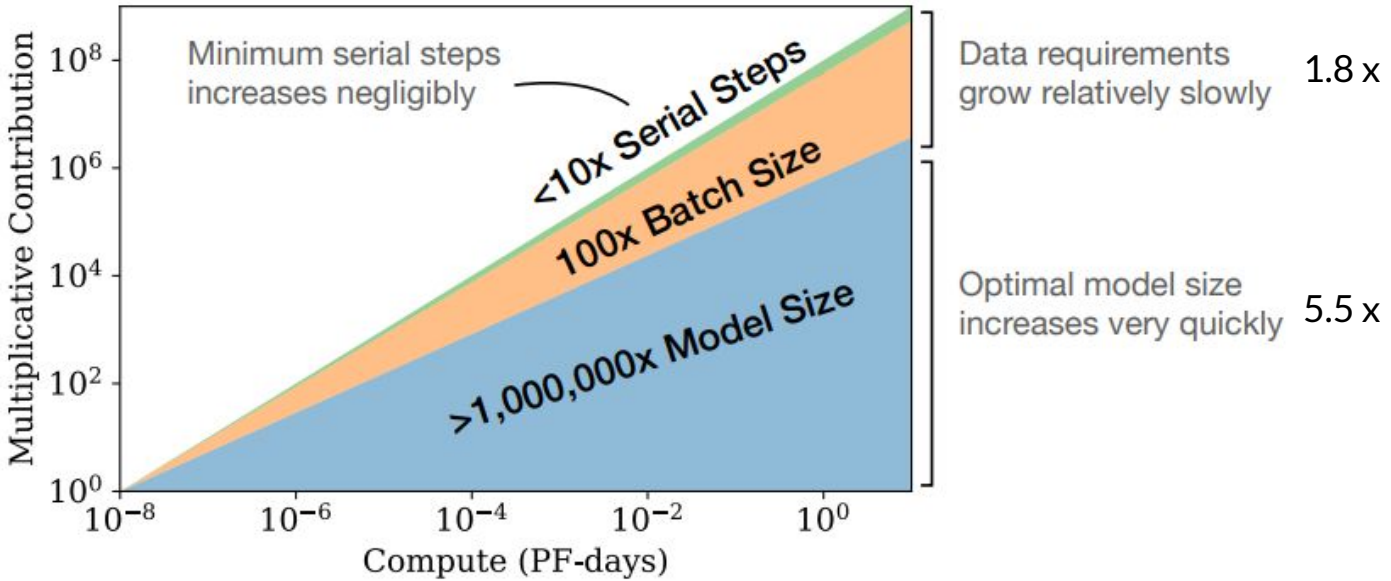

Source: Scaling Laws for Neural Language Models (Kaplan et. al)

# Improving model training

# Chinchilla

- Hypothesis: A smaller model trained on more data will perform better.
- Tested on 400 language models, 70 million to over 16 billion parameters.
- Datasets from 5 to 500 billion tokens.
- Chinchilla - 70B and 1.4T training tokens.
- Outperforms all previous models
- Less compute for fine-tuning and inference

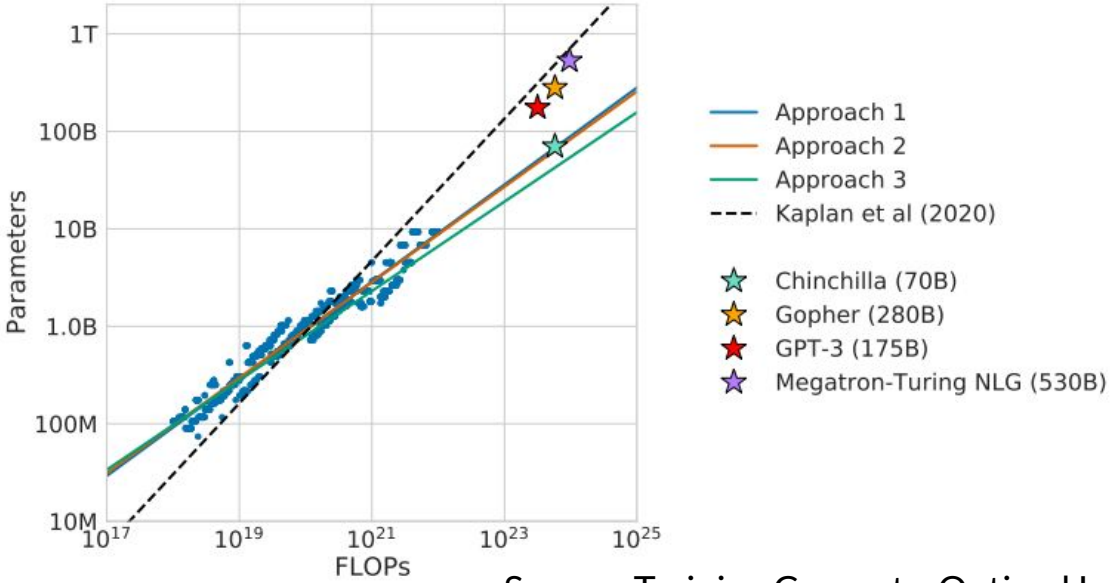
# Scaling laws




Source: Scaling Laws for Neural Language Models (Kaplan et. al)

Recommendation from Chinchilla paper:

For a 10 fold increase in computational budget, the model size and the number of training tokens should be scaled in equal proportions.




Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# Training Tokens

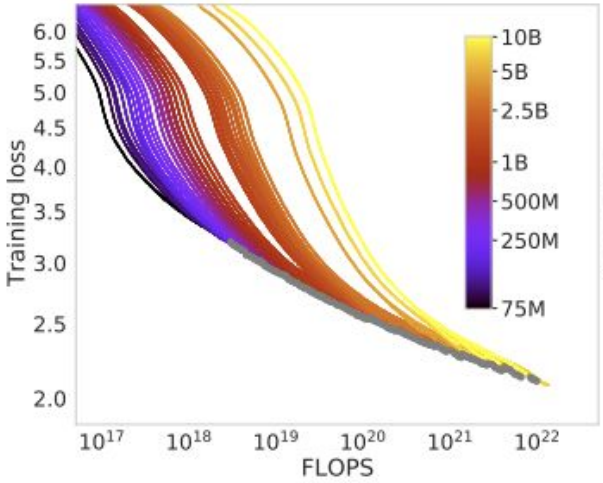
| Model | Size (# Parameters) | Training Tokens |
|---|---|---|
| LaMDA (Thoppilan et al., 2022) | 137 Billion | 168 Billion |
| GPT-3 (Brown et al., 2020) | 175 Billion | 300 Billion |
| Jurassic (Lieber et al., 2021) | 178 Billion | 300 Billion |
| *Gopher* (Rae et al., 2021) | 280 Billion | 300 Billion |
| MT-NLG 530B (Smith et al., 2022) | 530 Billion | 270 Billion |
| *Chinchilla* | 70 Billion | 1.4 Trillion |

Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# DeepMind team wanted to answer this question

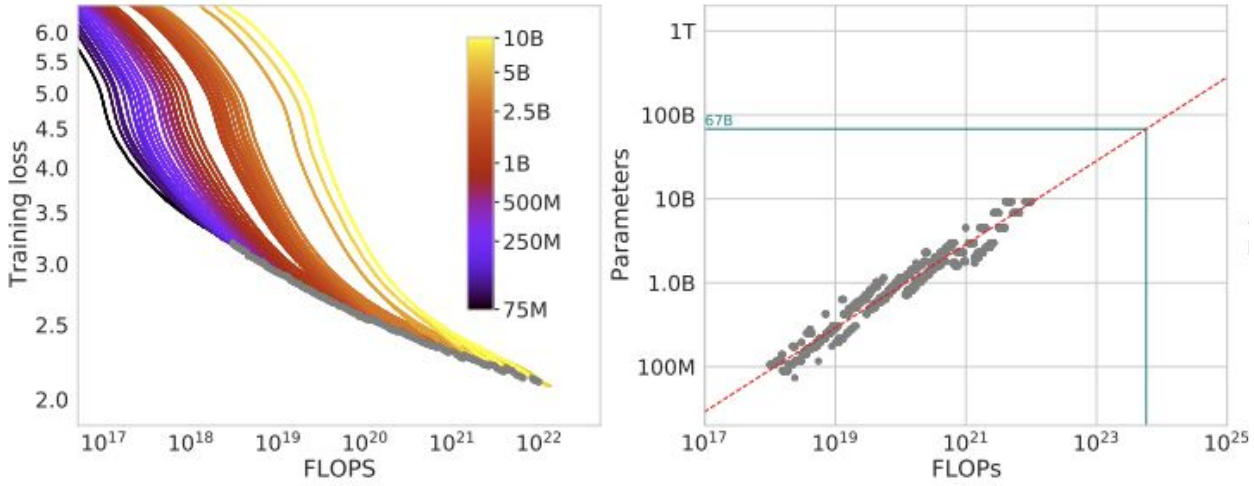
Given a fixed FLOPs budget, how should one trade-off model size and the number of training tokens?

# Fix the model size and vary number of training tokens



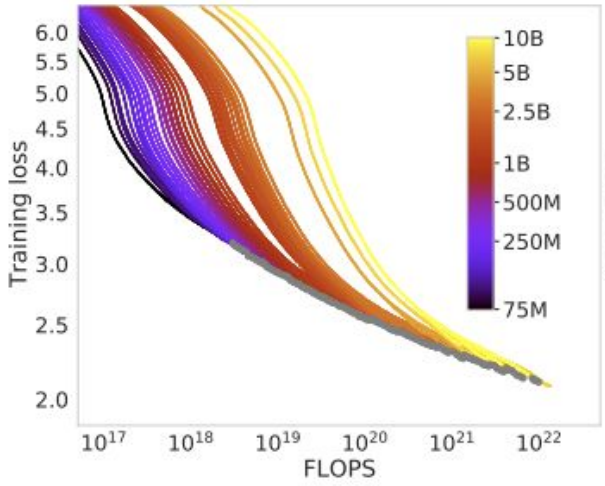

Source: Training Compute-Optimal Large Language Models (Hoffman et al)

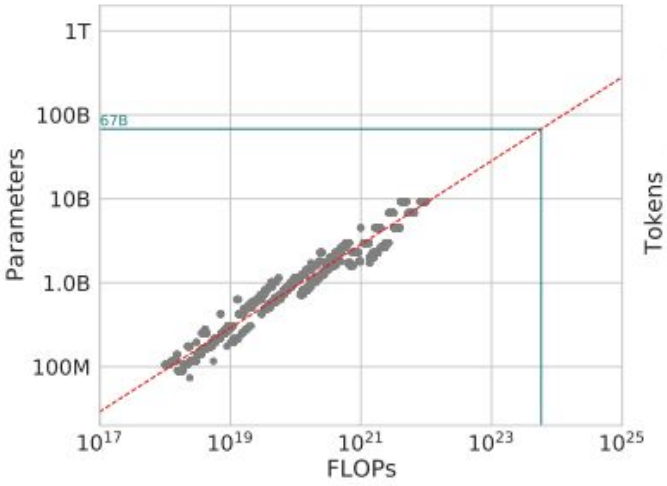
# Fix the model size and vary number of training tokens

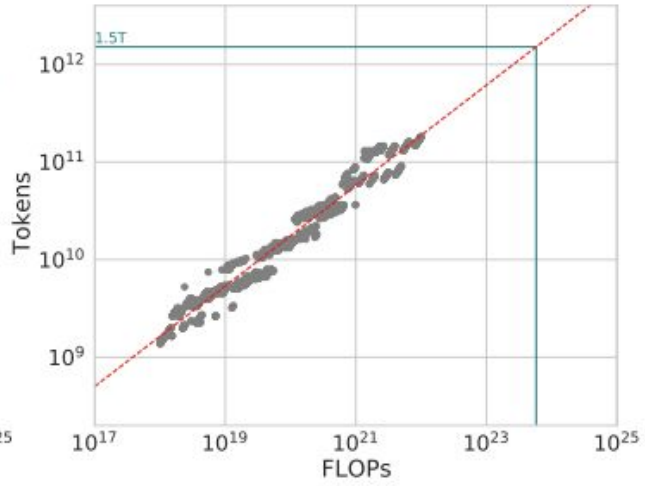



Source: Training Compute-Optimal Large Language Models (Hoffman et al)

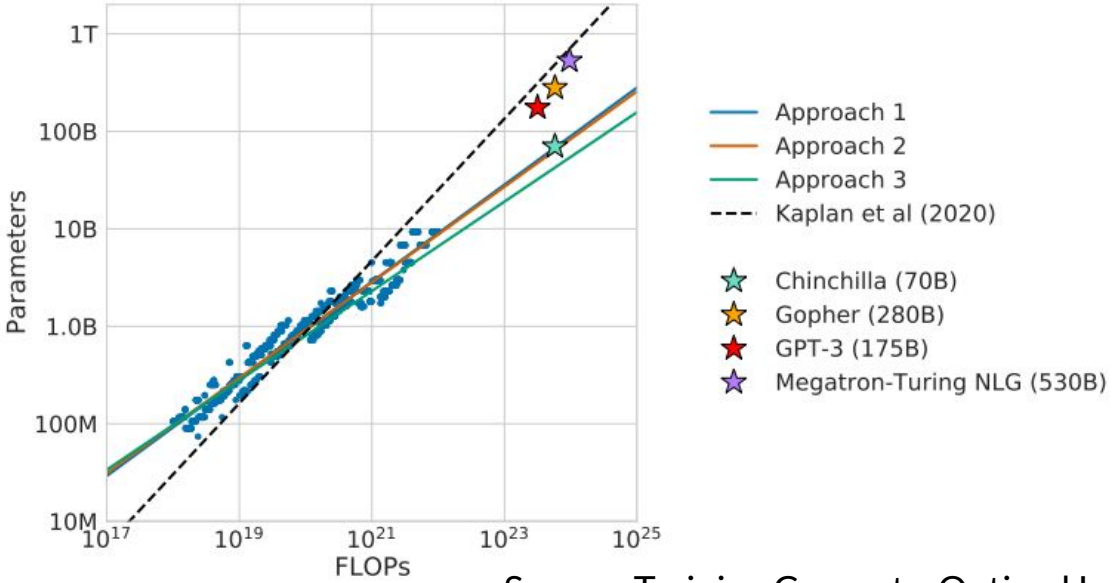
# Fix the model size and vary number of training tokens



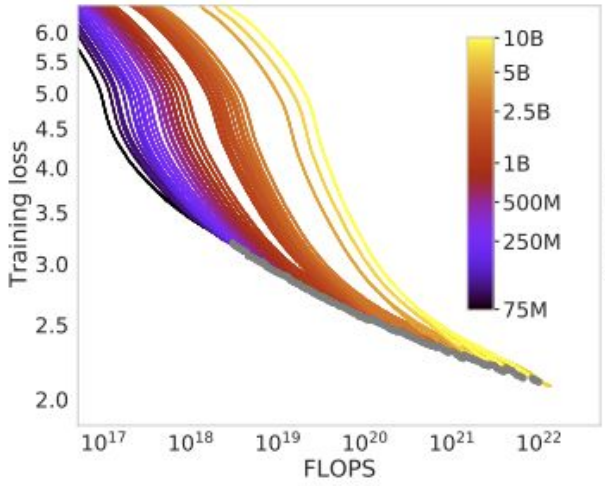



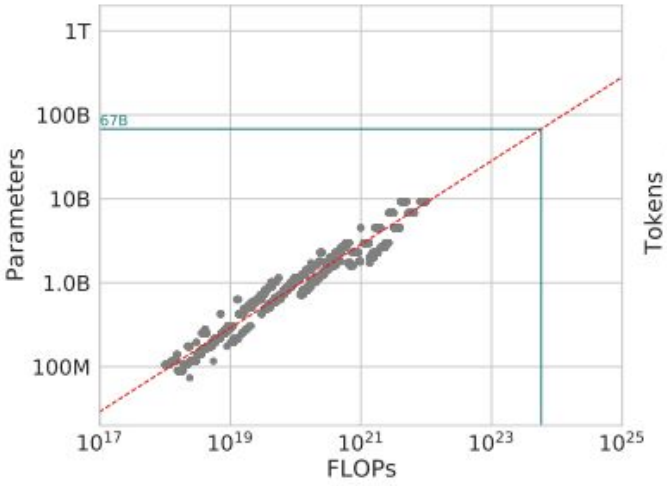

Source: Training Compute-Optimal Large Language Models (Hoffman et al)

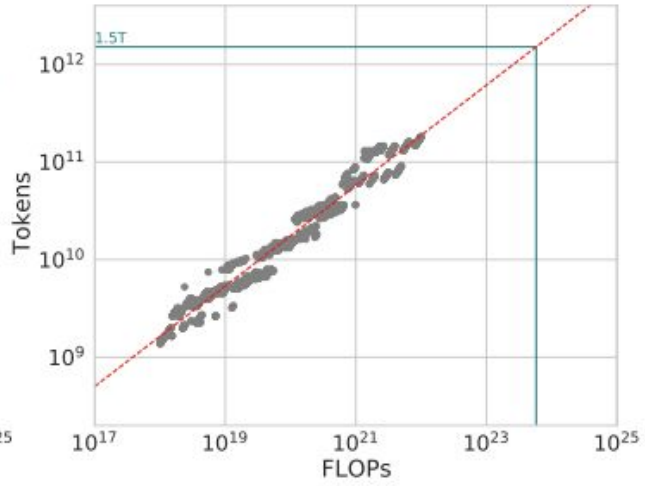



Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# Fix the model size and vary number of training tokens

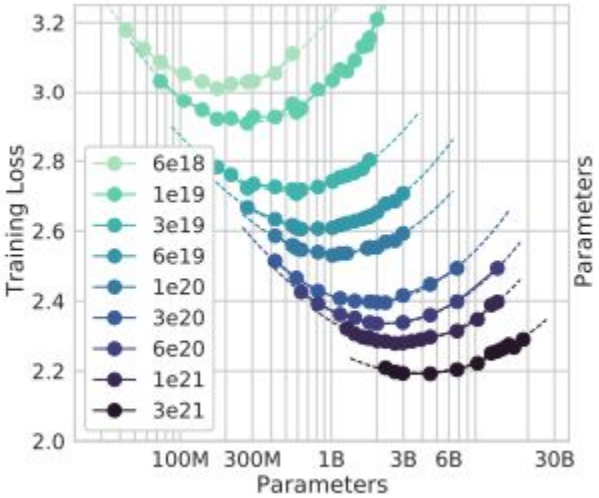


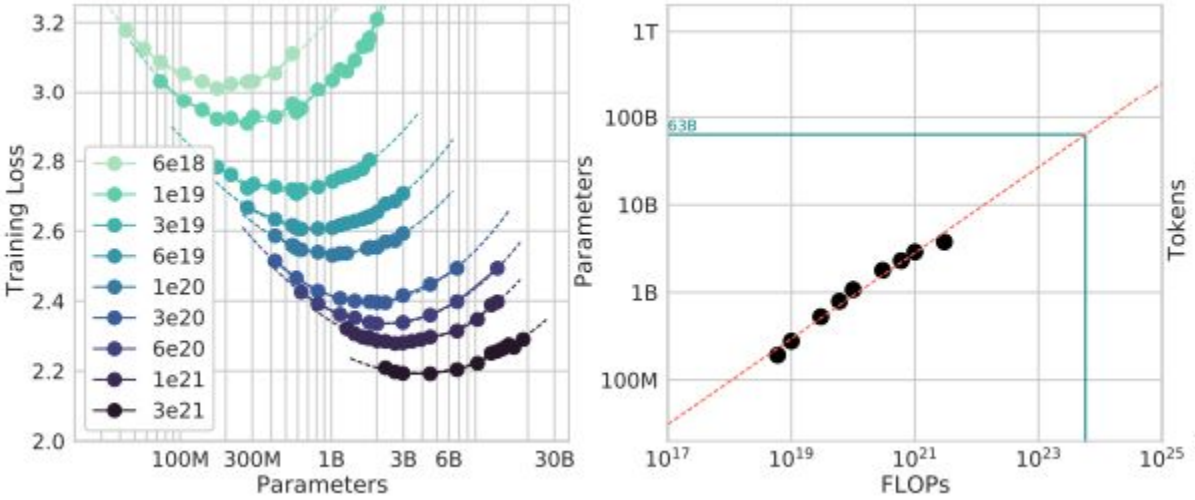




Source: Training Compute-Optimal Large Language Models (Hoffman et al)

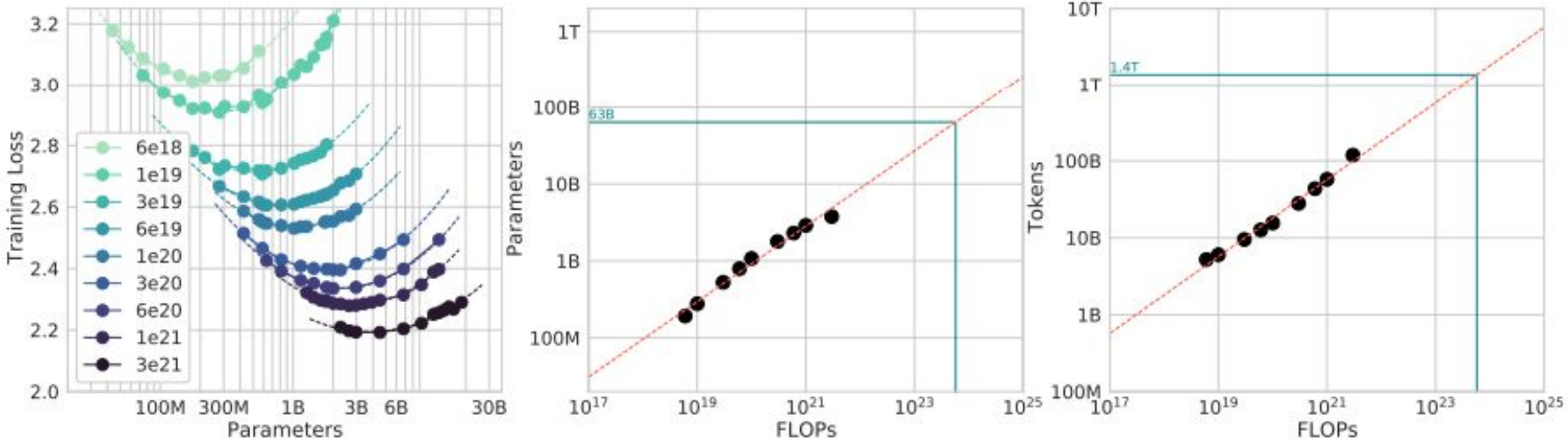
# For a given FLOP budget, what is the optimal parameter count?




Source: Training Compute-Optimal Large Language Models (Hoffman et al)




Source: Training Compute-Optimal Large Language Models (Hoffman et al)

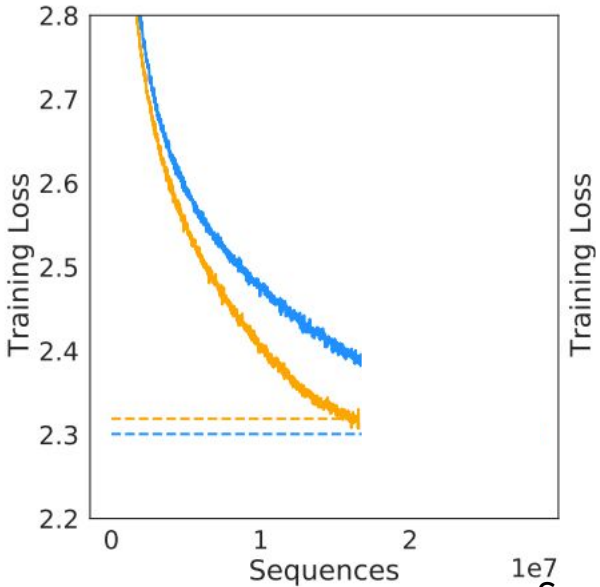



Source: Training Compute-Optimal Large Language Models (Hoffman et al)

| Parameters | FLOPs | FLOPs (in Gopher unit) | Tokens |
|---|---|---|---|
| 67 Billion | 5.76E+23 | 1 | 1.5 Trillion |
| 175 Billion | 3.85E+24 | 6.7 | 3.7 Trillion |
| 280 Billion | 9.90E+24 | 17.2 | 5.9. Trillion |

Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# Comparing with Scaling Laws

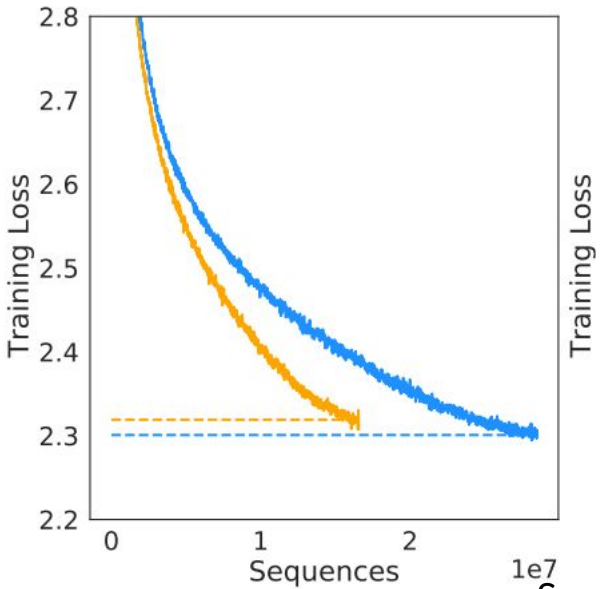



Chinchilla

Scaling laws

Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# Comparing with Scaling Laws

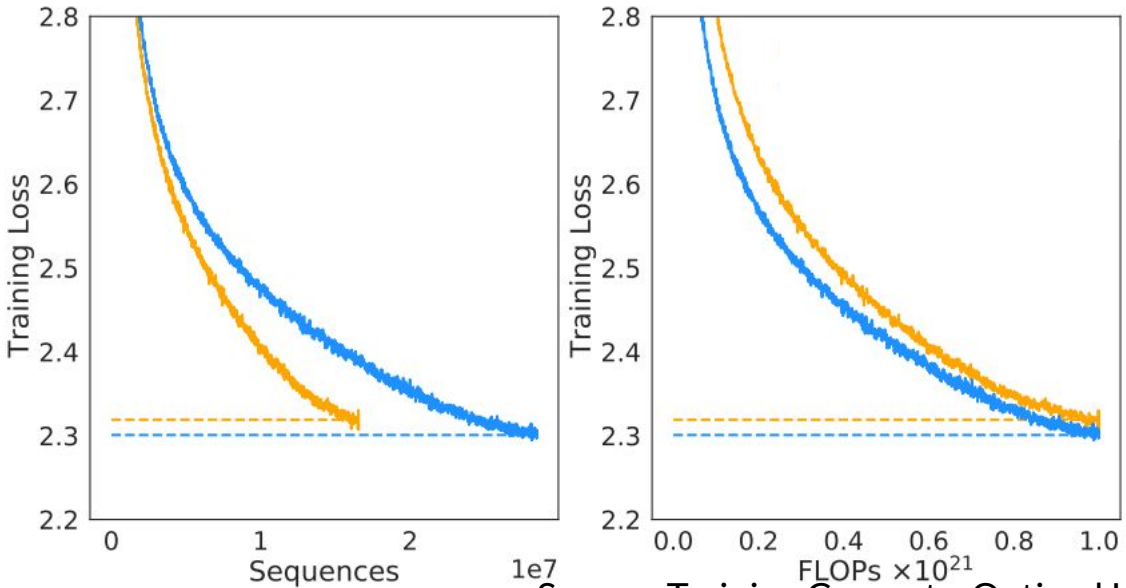



Chinchilla

Scaling laws

Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# Comparing with Scaling Laws




Source: Training Compute-Optimal Large Language Models (Hoffman et al)

# What are the benchmarks?

# MMLU: Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s²
(B) more than 9.8 m/s²
(C) less than 9.8 m/s²
(D) Cannot say unless the speed of throw is given.

# MMLU: Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

(A) 9.8 m/s² ✔

(B) more than 9.8 m/s² ✘

(C) less than 9.8 m/s² ✘

(D) Cannot say unless the speed of throw is given. ✘

# MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained.
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
(C) monopoly firms do not engage in significant research and development.
(D) consumer surplus is lost with higher prices and lower levels of output.

# MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

# MMLU: Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk
(B) Branch of the external carotid artery
(C) Branch of the thyrocervical trunk
(D) Tributary of the internal jugular vein

# MMLU: Medicine

A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk ✗
(B) Branch of the external carotid artery ✗
(C) Branch of the thyrocervical trunk ✓
(D) Tributary of the internal jugular vein ✗

# MMLU-Pro (released 2024)

A small percentage of MMLU is wrong/ambigious. MMLU

MMLU is only 4 choices - MMLU-Pro is 10

https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro/viewer

# Benchmarks are always getting deprecated

Why SWE-bench Verified no longer measures frontier coding capabilities

# Current trends

# Open vs. Closed models

# OSS licensing

**Non-commercial license**

- Creative Commons CC BY-NC-SA 4.0

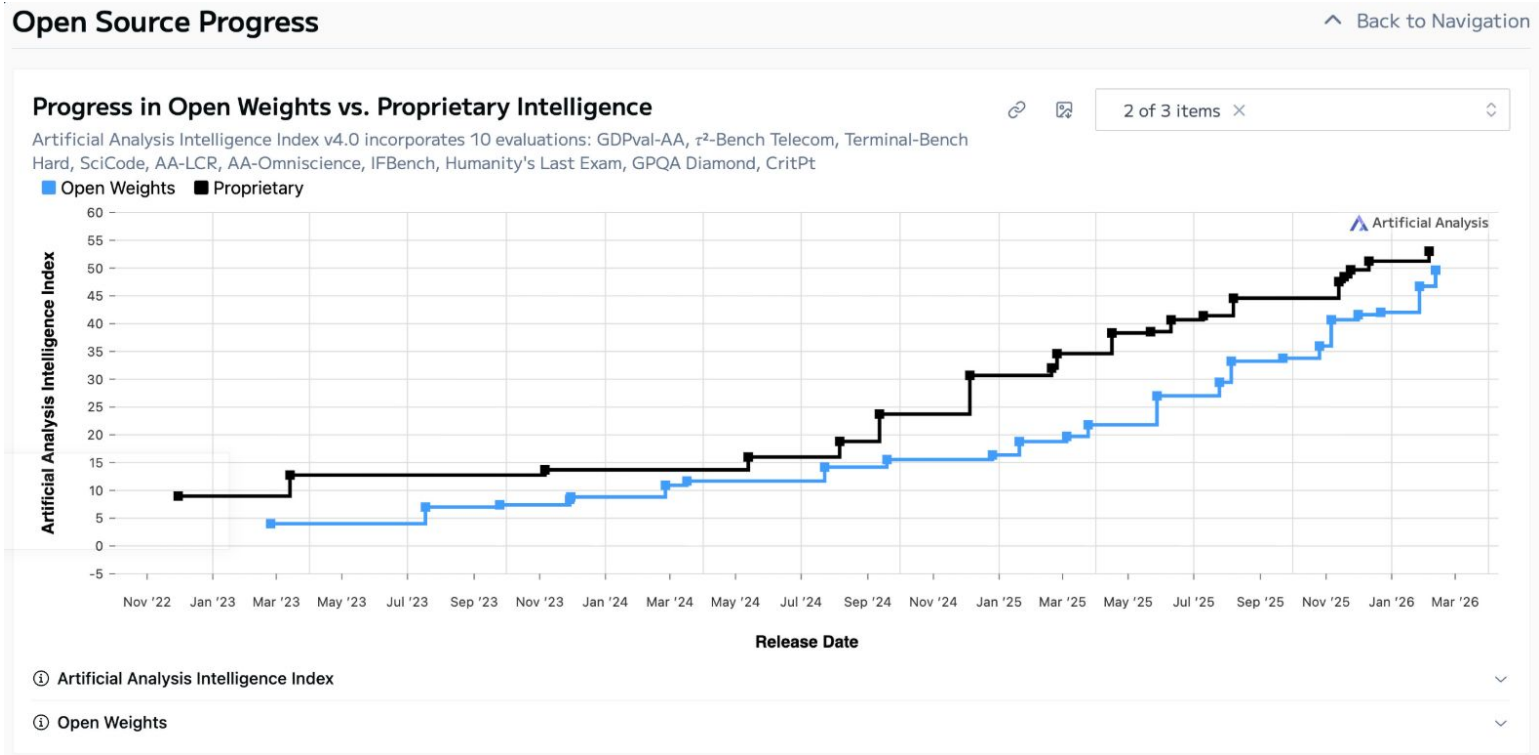
**Restricted license**

- CC BY-SA- 3.0 restrictions on commercial use

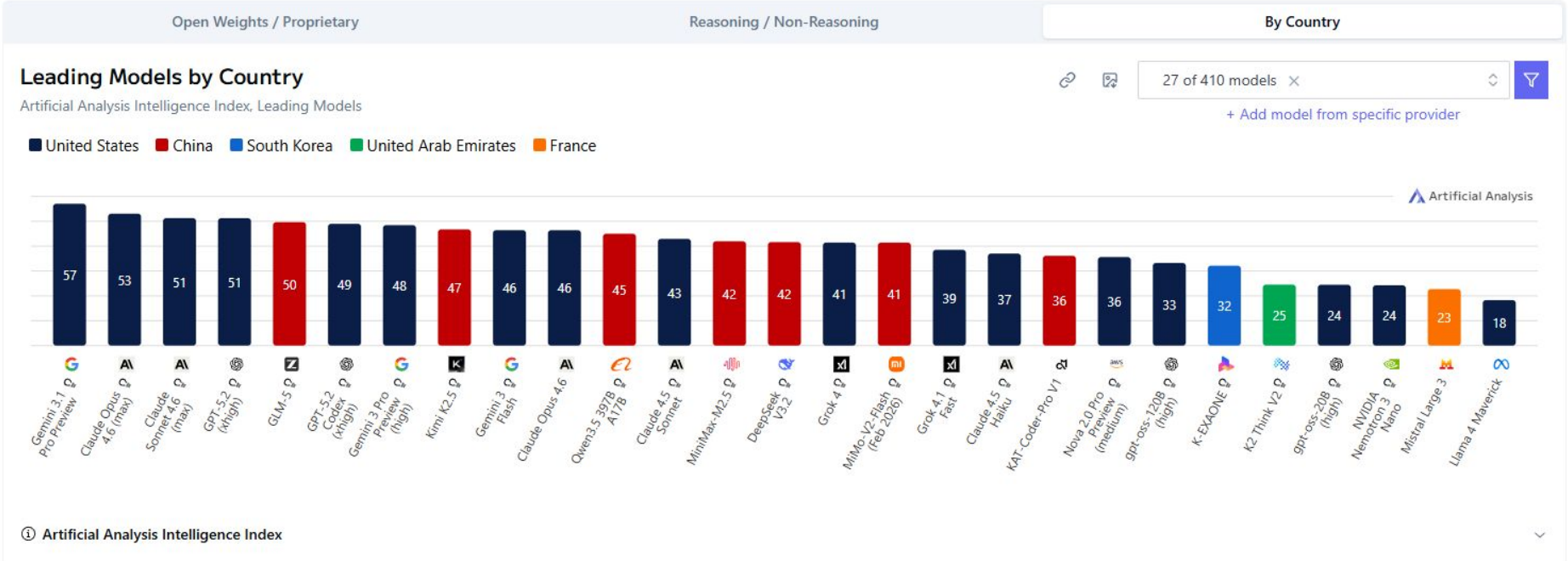
**Permissive license**

- Apache 2.0

# OSS licensing



# Open Weight models by Country



# NVIDIA

# Why would  NVIDIA want to support Open Models?

# NVIDIA – Nemotron

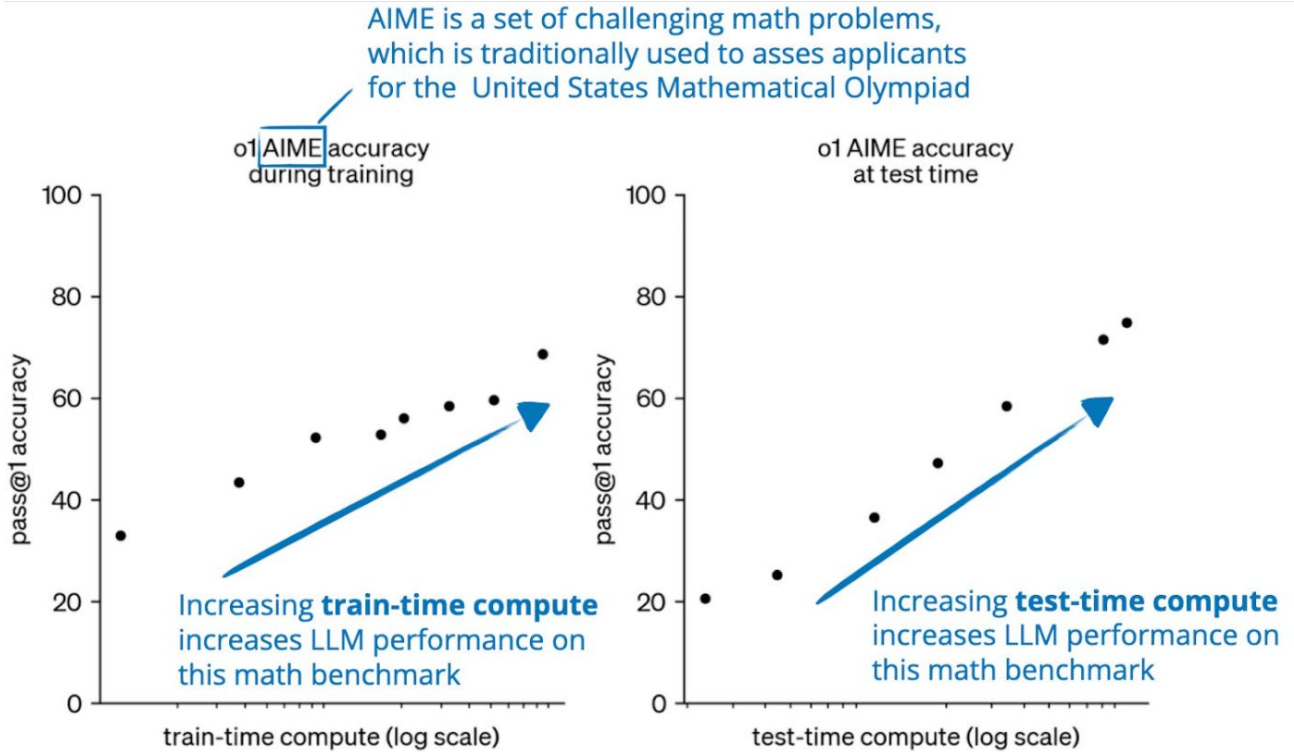
- Family of open models (and related assets) aimed at building specialized AI agents.
- Models are transparent and “open”: open weights, training data, and training recipes/technical
- Range of modalities/capabilities:
  - Vision-language (visual understanding),
  - RAG (retrieval models like embed/rerank),
  - speech (ASR/TTS/NMT), and
  - safety model

# NVIDIA – Nemotron

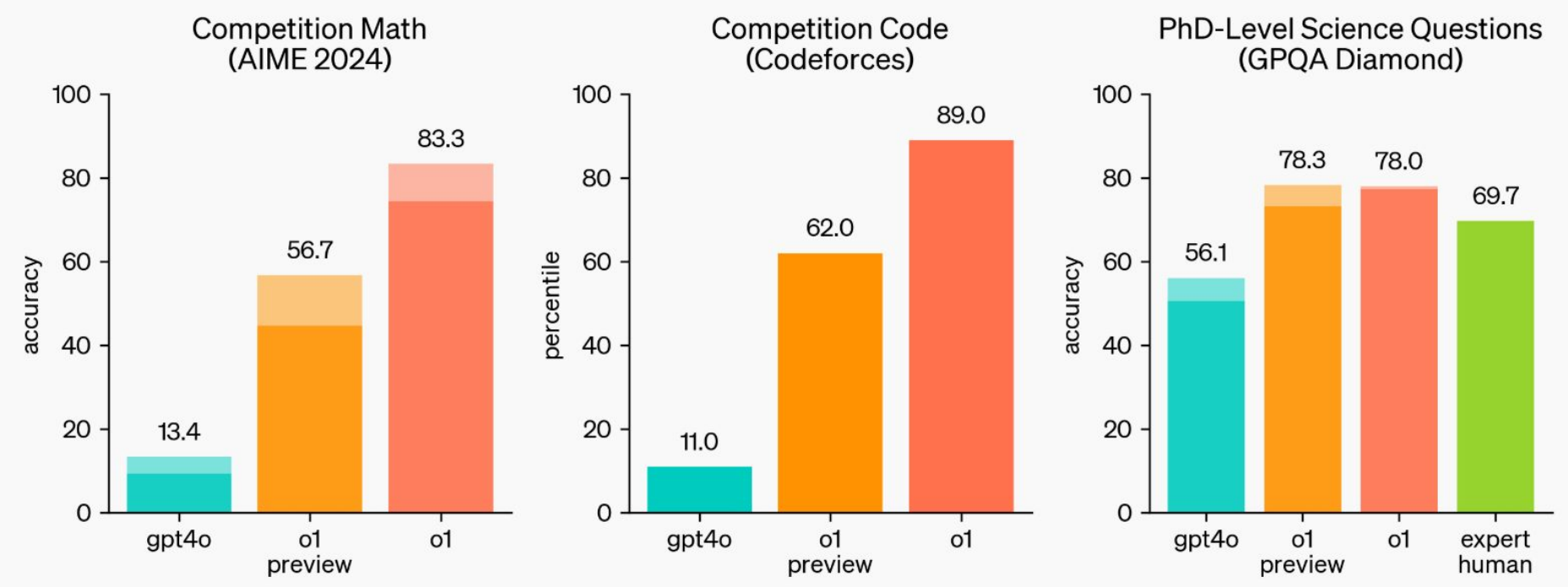
- Open datasets for agentic AI (pre/post-training, personas, safety, RL, RAG), commercially usable and spanning very large scale (e.g., 10T+ tokens and millions of SFT samples).
- Reasoning tiers optimized for different needs:
  - Nano (cost-efficient, targeted agentic tasks)
  - Super (high accuracy for multi-agent reasoning)
  - Ultra (highest reasoning accuracy, large deployments)

# Reasoning models

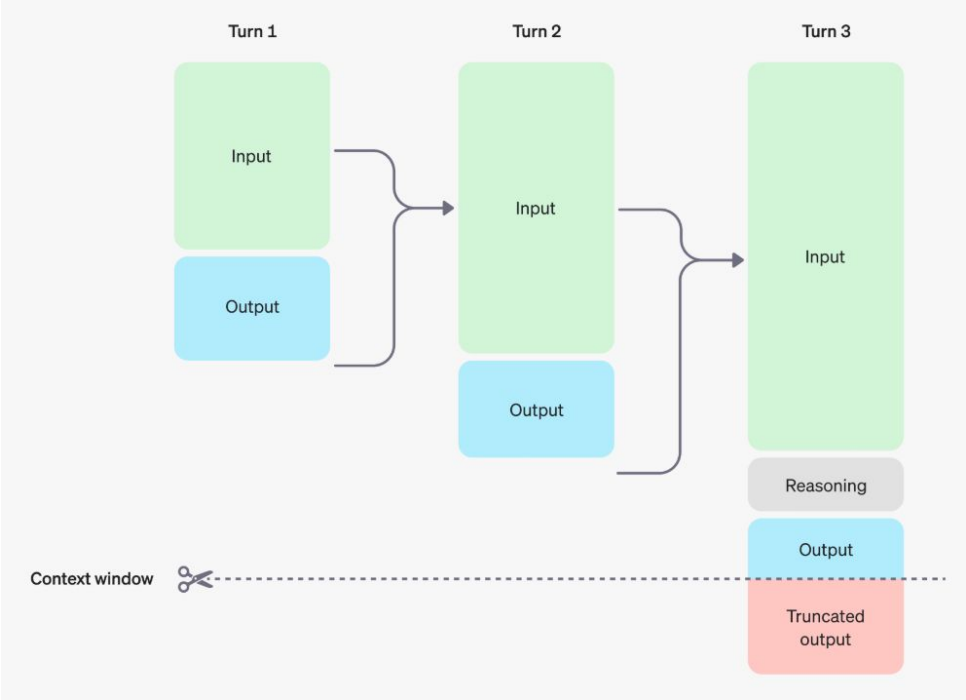
# What are reasoning models



# What are reasoning models



# What are reasoning models



# Reasoning models

Find the connection between the following four words:

**Incognito**

**Unbeknownst**

**Nonchalant**

**Misnomer**

# Comparing Reasoning models to non-reasoning models

**Assistant**

The connection between the words "Incognito," "Unbeknownst," "Nonchalant," and "Misnomer" lies in the theme of disguise, misunderstanding, or things not being what they seem:

1. **Incognito** - This term refers to a state of disguise or having one's identity concealed. It involves presenting oneself in a way that is different from one's true identity.

2. **Unbeknownst** - This word means something happening without the knowledge of someone, implying a lack of awareness or understanding, often about something significant or relevant.

3. **Nonchalant** - This describes an attitude of indifference or casualness, suggesting a disguise or masking of true emotions or concerns, often giving an impression of not caring or being unaffected.

4. **Misnomer** - A misnomer is a name or term that suggests an incorrect or misleading understanding of its subject, often disguising its true nature or identity due to being improperly named.

Together, these words share a commonality of disguise, whether it be in identity, awareness, emotions, or names, each relating to a theme of misunderstood or concealed truths.

4,806ms ↑31t ↓233t Request ID

# Reasoning models

These words are intriguing because they suggest that there should be a positive counterpart without the prefix, but in practice, only the "negative" form is used. This phenomenon highlights interesting quirks in the development and adoption of words in the English language.

**Answer:** Each is a “negative” word whose positive form is unused—they’re unpaired words missing their opposites.

# Reasoning/non-reasoning models

# Mixture of Experts

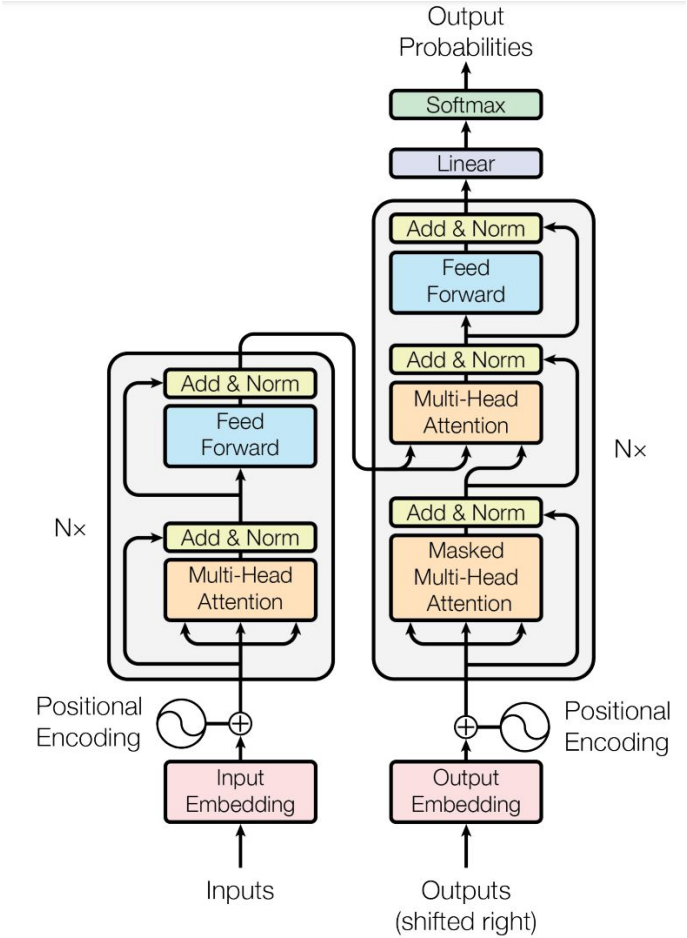
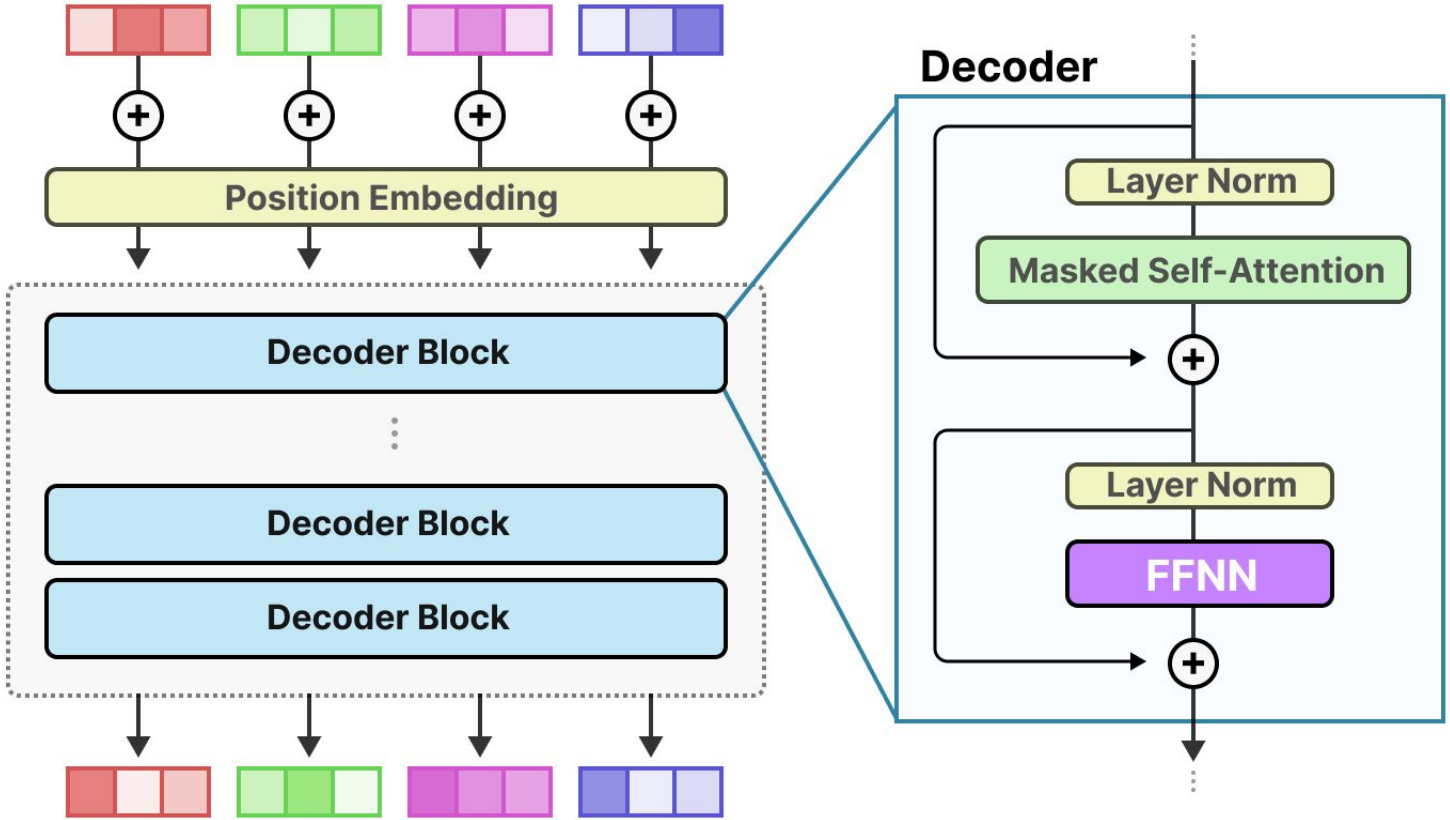
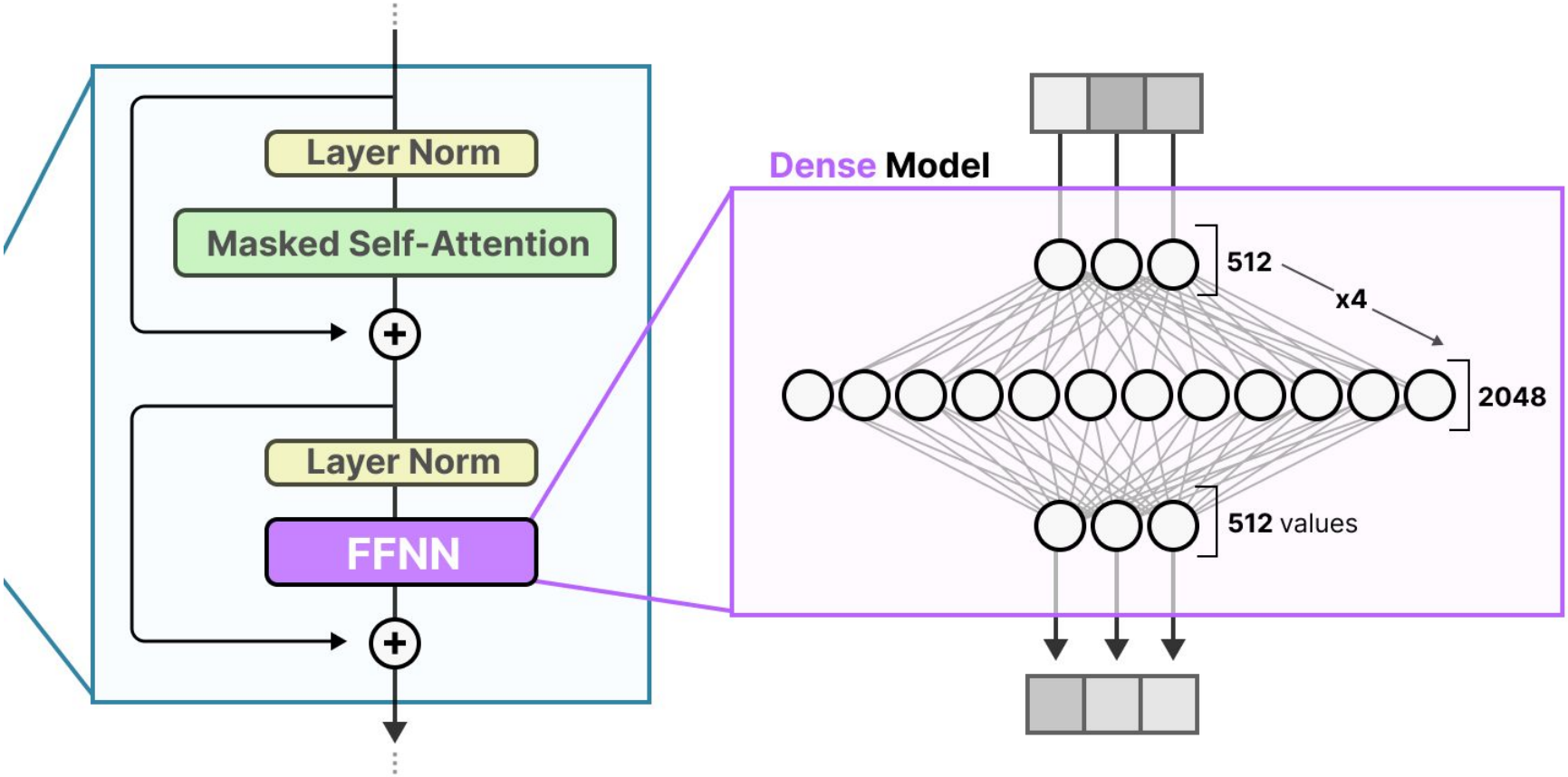



Figure 1: The Transformer - model architecture.

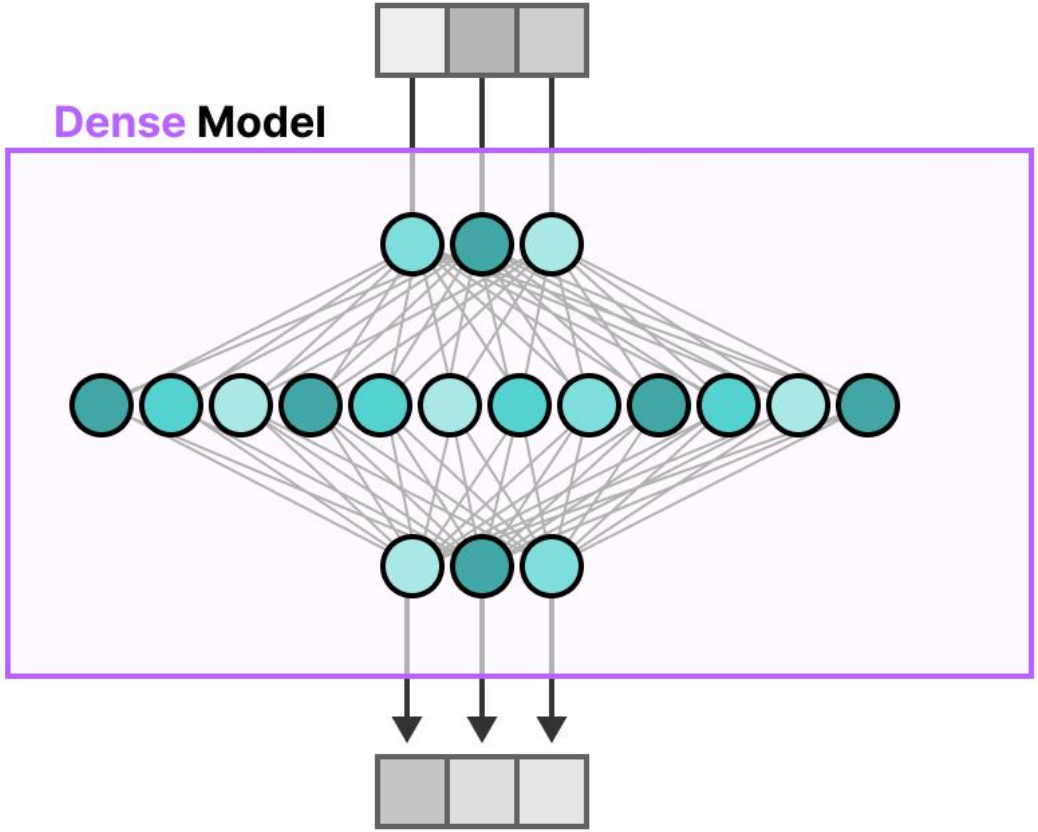
Attention is all you need



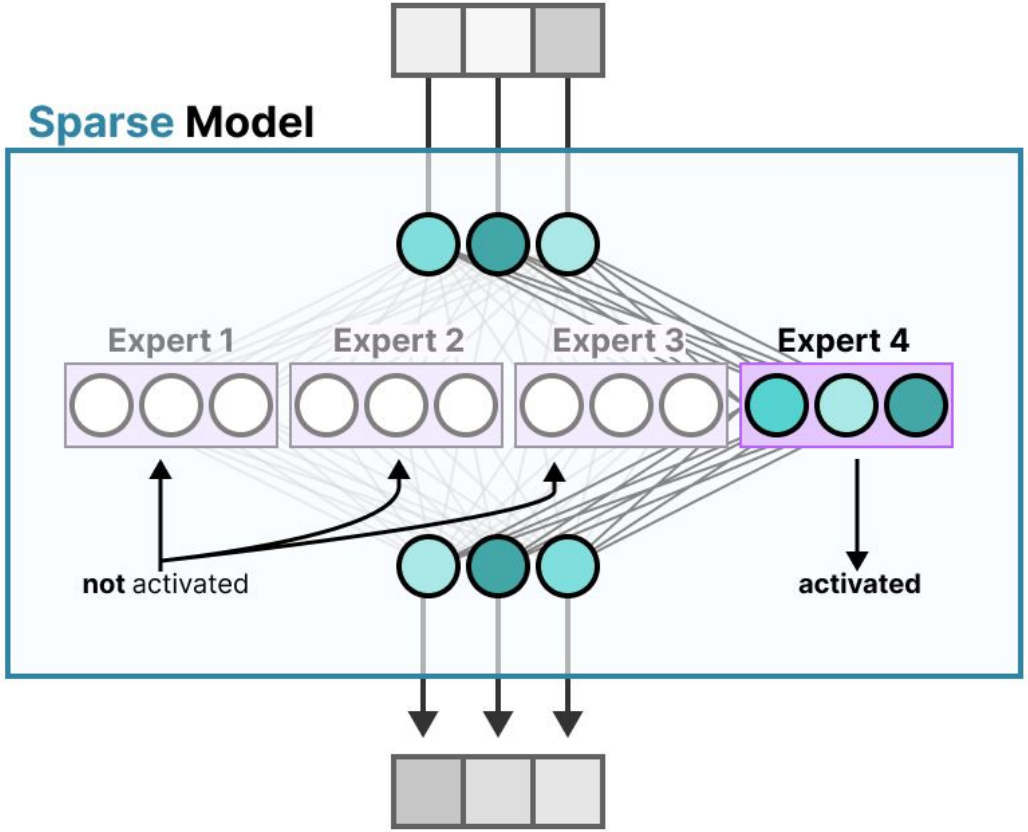

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts



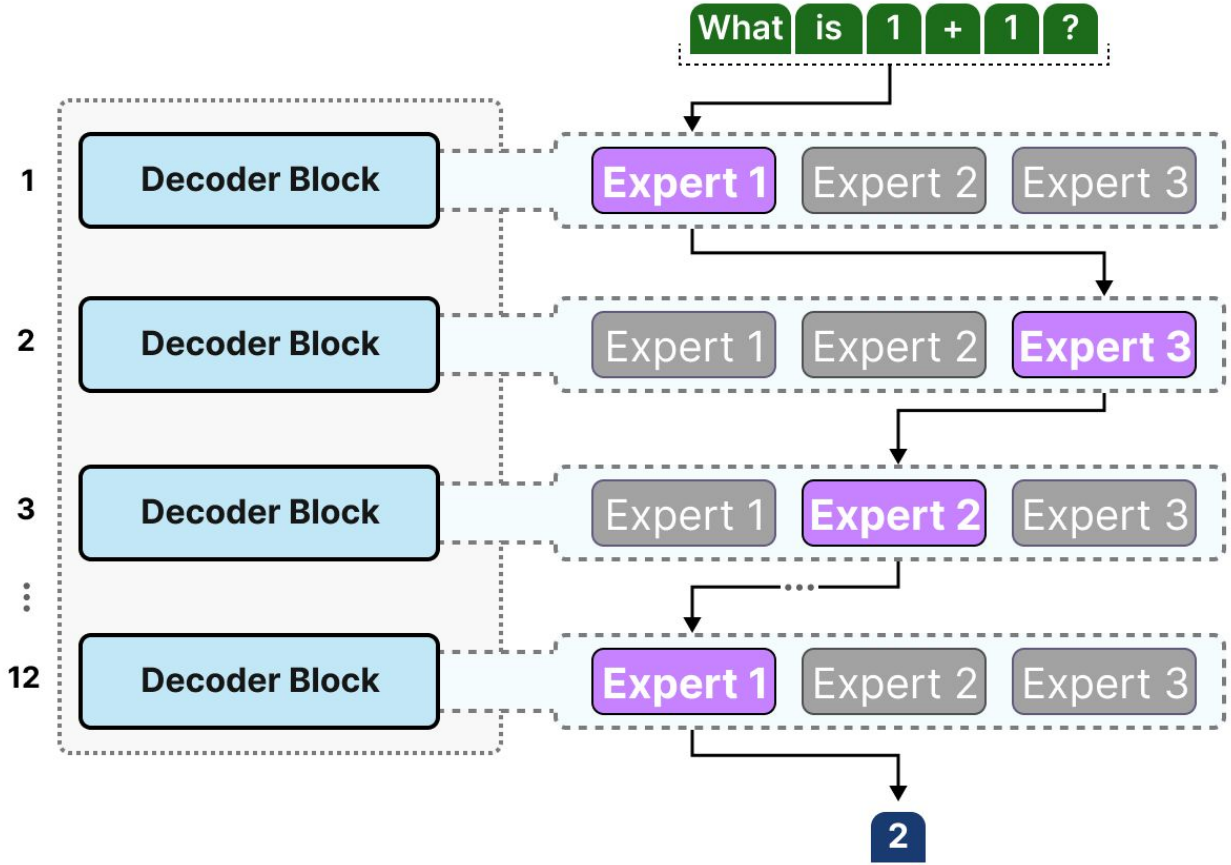

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts



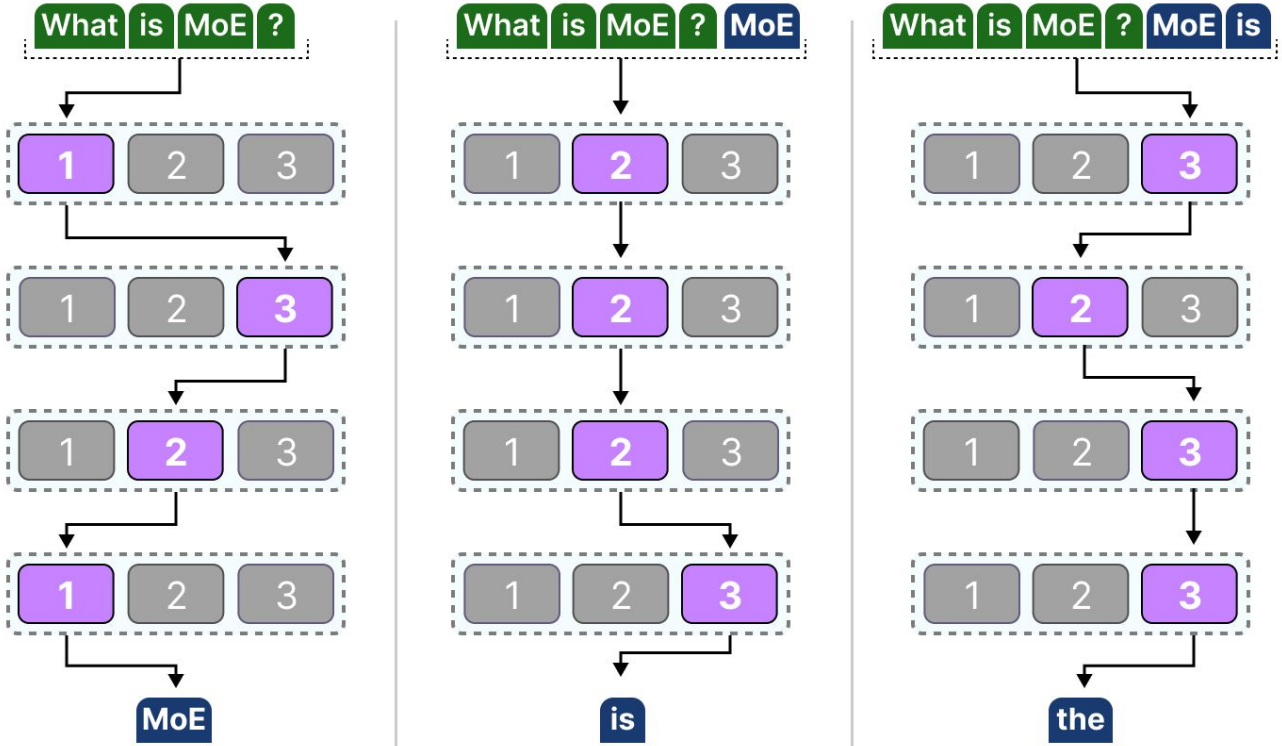

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts




https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts




https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts




https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mixture-of-experts

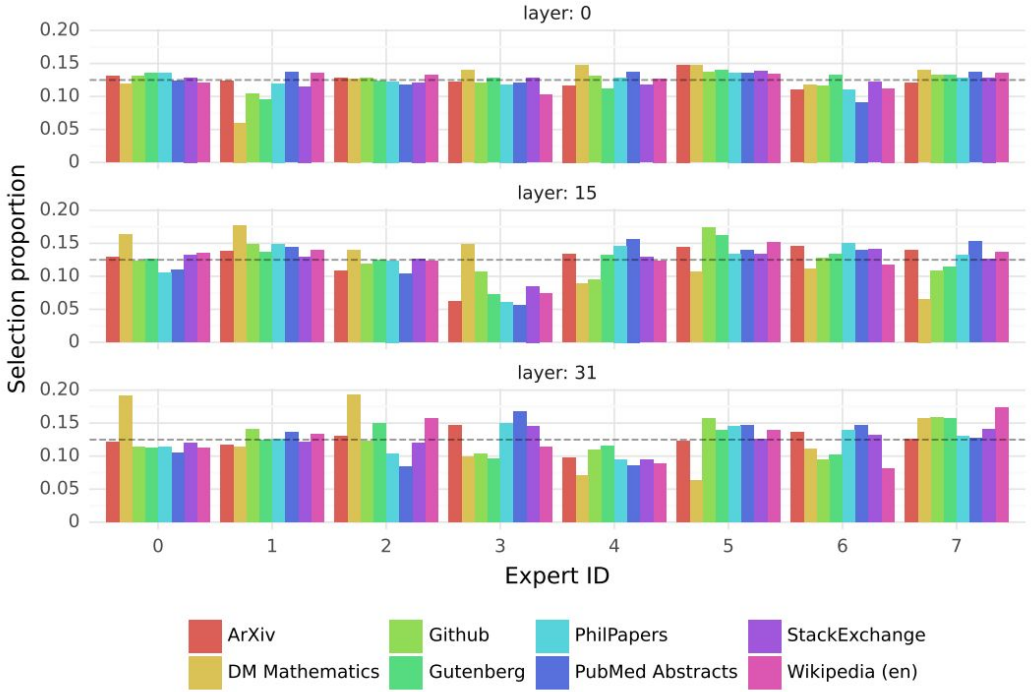



**Figure 7: Proportion of tokens assigned to each expert on different domains from The Pile dataset for layers 0, 15, and 31.** The gray dashed vertical line marks $1/8$, i.e. the proportion expected with uniform sampling. Here, we consider experts that are either selected as a first or second choice by the router. A breakdown of the proportion of assignments done in each case cane be seen in Figure 9 in the Appendix.

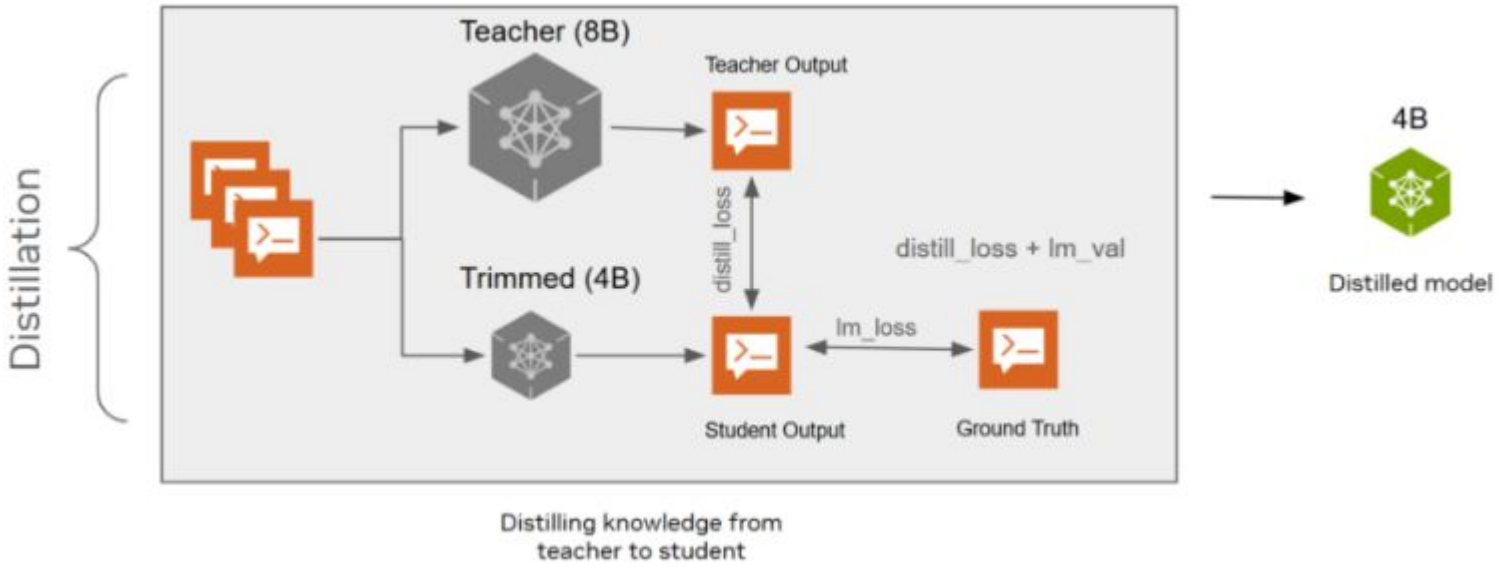
https://arxiv.org/pdf/2401.04088


**Figure 8: Text samples where each token is colored with the first expert choice.** The selection of experts appears to be more aligned with the syntax rather than the domain, especially at the initial and final layers.

https://arxiv.org/pdf/2401.04088

# Model distillation

# Model Distillation




Distilling knowledge from teacher to student

# Model Distillation

13th Feb 2026

We have identified industrial-scale campaigns by three AI laboratories—DeepSeek, Moonshot, and MiniMax—to illicitly extract Claude's capabilities to improve their own models. These labs generated over 16 million exchanges with Claude through approximately 24,000 fraudulent accounts, in violation of our terms of service and regional access restrictions.

These labs used a technique called "distillation," which involves training a less capable model on the outputs of a stronger one. Distillation is a widely used and legitimate training method. For example, frontier AI labs routinely distill their own models to create smaller, cheaper versions for their customers. But distillation can also be used for illicit purposes: competitors can use it to acquire powerful capabilities from other labs in a fraction of the time, and at a fraction of the cost, that it would take to develop them independently.

# Model Distillation

- Anthropic says it detected **industrial-scale “distillation attacks”** targeting Claude by **three AI labs: DeepSeek, Moonshot, and MiniMax**.
- The campaigns used **~24,000 fraudulent accounts** and generated **over 16 million exchanges** with Claude, violating terms of service and regional restrictions.
- **Distillation itself is a legitimate ML technique** (used to make smaller/cheaper models), but becomes a problem when competitors use it **illicitly to copy capabilities** quickly and cheaply.

# Model Distillation

**DeepSeek**

*Scale: Over 150,000 exchanges*

The operation targeted:

- Reasoning capabilities across diverse tasks
- Rubric-based grading tasks that made Claude function as a reward model for reinforcement learning
- Creating censorship-safe alternatives to policy sensitive queries

# Model Distillation

**Moonshot AI**

*Scale: Over 3.4 million exchanges*

The operation targeted:

- Agentic reasoning and tool use
- Coding and data analysis
- Computer-use agent development
- Computer vision

# Model Distillation

**MiniMax**

*Scale: Over 13 million exchanges*

The operation targeted:

- Agentic coding
- Tool use and orchestration

# Why distillation?

Distillation is a shortcut to more compute for anyone.

Training requires a big GPU cluster, so significant capital expenditure (high risk)

APIs are pay-as-you-go, so operational expenditure (low risk)

# How are the attacks carried out?

- Claude access is restricted in China (and some related entities).
- Labs bypass this via proxy resellers.
- Proxies use “hydra clusters” of fake accounts.
- Traffic is spread across APIs + cloud platforms.
- Ban one account, another replaces it.
- One network ran 20,000+ fake accounts.
- Distillation traffic is mixed with normal requests.
- Labs send massive, targeted prompt batches.
- Outputs become training data or RL tasks.

- Distillation shows up as patterns, not single prompts.
- Signals: high volume, narrow focus, repetitive prompts.

> *You are an expert data analyst combining statistical rigor with deep domain knowledge. Your goal is to deliver data-driven insights — not summaries or visualizations — grounded in real data and supported by complete and transparent reasoning.*

# Fine-tuning

# Why fine-tuning?

- Improve task performance
- Adapt to a specific domain
- Enforce a certain output style
- Lower inference cost and latency

# Fine-tuning Lab

https://colab.research.google.com/drive/1qLgiL0kQ008PX5NCR29cv3T7CmmL9e_y?usp=sharing


O'REILLY®